Adversarial Risk Analysis

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
Via Alfonso Corti 12, I-20133, Milano, Italy, European Union

fabrizio@mi.imati.cnr.it

www.mi.imati.cnr.it/fabrizio

OUTLINE OF THE COURSE

- Introduction to Bayesian Statistics
- Introduction to Adversarial Risk Analysis
- Discrete Simultaneous Games and Modelling Opponents
- Example: Auctions
- Sequential Games
- Example: Somali Pirates
- My works
  - Adversarial Hypothesis Testing
  - Batch Acceptance
  - Classification
  - Software Release

ALL BAYESIANS IN DAILY LIFE?

Interest in Milano or not?

- Prior knowledge
  - What is Milano? City, cookie, meat, car?
  - Where is the city of Milano?
  - Fashion and football
- Data collection
  - Book on snorkeling activities
  - Tour operator catalogue
  - City of Milano official website

ALL BAYESIANS IN DAILY LIFE?

- Posterior knowledge
  - No snorkeling: closest beach at 150 kms!
  - Probably no tour found in the catalogue
  - Leonardo's Last Supper; Michelangelo, Raffaello, Mantegna, etc.; Duomo (cathedral); Sforza Castle; Canals (Navigli) and nightlife; Via Sarpi (Chinatown); etc.
- Forecast:
  - Will I enjoy Milano or not?
  - Cost and time to get there
- Decision: To go or not to go?
  - Interest in the place
  - Distance and cost for travel, lodging and meals
  - Italian language (but English understood by many)

BAYES THEOREM

- Patient subject to medical diagnostic test (P or N) for a disease D
- *Sensitivity* .95, i.e. $\mathbb{P}(P|D) = .95$
- *Specificity* .9, i.e. $\mathbb{P}(P^C|D^C) = P(N|D^C) = .9$
- Physician's belief on patient having the disease 1%, i.e. $\mathbb{P}(D) = .01$
  - Knowledge about **that** patient
  - Knowledge about people with similar characteristics (age, gender, etc.)
  - Knowledge about the population in an area
  - Other sources of knowledge or uninformative guess
- Positive test $\Rightarrow \mathbb{P}(D|P)$?

BAYES THEOREM

$$\begin{aligned}\mathbb{P}(D|P) &= \frac{\mathbb{P}(D \cap P)}{\mathbb{P}(P)} = \frac{\mathbb{P}(P|D)\mathbb{P}(D)}{\mathbb{P}(P|D)\mathbb{P}(D) + \mathbb{P}(P|D^C)\mathbb{P}(D^C)} \\ &= \frac{.95 \cdot .01}{.95 \cdot .01 + .1 \cdot .99} = .0875\end{aligned}$$

Positive test updates belief on patient having the disease:
from 1% to 8.75%

Prior opinion updated into *posterior* one

If $\mathbb{P}(D) = .1 \Rightarrow \mathbb{P}(D|P) = .5135$
If $\mathbb{P}(D) = .2 \Rightarrow \mathbb{P}(D|P) = .7037$

BAYES THEOREM

- Partition $\{A_1, \ldots, A_n\}$ of Ω and $B \subset \Omega : \mathbb{P}(B) > 0$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)P(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)P(A_j)}$$

- X r.v. with density $f(x|\lambda)$, prior $\pi(\lambda)$

$$\Rightarrow \text{posterior } \pi(\lambda|x) = \frac{f(x|\lambda)\pi(\lambda)}{\int f(x|\omega)\pi(\omega)d\omega}$$

A SHORT HISTORY OF BAYESIAN STATISTICS

- Bayesian statistics strongly relies on the use of Bayes Theorem
- The idea of Bayes Theorem goes back to James Bernoulli in 1713 but there was no mathematical structure yet
- Reverend Thomas Bayes died in 1761
- Richard Price, Bayes's friend, published Bayes's paper on inverse probability in 1763, which was about binomial data and uniform prior
- In 1774 Laplace gave more general results, probably unaware of Bayes's work
- Jeffreys "rediscovered" Bayes's work in 1939
- Bruno de Finetti and Jimmy Savage set the foundations of the Bayesian approach
- In early 90's Metropolis simulation method was "ridiscovered" by Gelfand and Smith
- Since then MCMC (Markov chain Monte Carlo) and other simulation methods were developed and Bayesian approach became very popular

ASSESSMENT OF PRIOR PROBABILITIES

Bayesian Statistics relies on subjective assessment of probabilities, but have a look at this example:

- T= person having a tumor in his/her life
- I= person having an infarction in his/her life
- Are these probability assessments right or not?
  1. $\mathbb{P}(T \bigcup I) = .2,\ \mathbb{P}(T) = .3,\ \mathbb{P}(I) = .05,\ \mathbb{P}(T \bigcap I) = .1$
  2. $\mathbb{P}(T \bigcup I) = .3,\ \mathbb{P}(T) = .2,\ \mathbb{P}(I) = .2,\ \mathbb{P}(T \bigcap I) = .15$
  3. $\mathbb{P}(T \bigcup I) = .3,\ \mathbb{P}(T) = .2,\ \mathbb{P}(I) = .2,\ \mathbb{P}(T \bigcap I) = .1$
- Assessments should comply with probability rules

ASSESSMENT OF PRIOR PROBABILITIES

- $P(A)$: Probability one of us was born on a given day, say May, 1st
- n people $\Rightarrow P(A) = 1 - (364/365)^n$
-

$$\begin{aligned} n = 10 &\Rightarrow P(A) = 0.027 \\ n = 50 &\Rightarrow P(A) = 0.128 \\ n = 100 &\Rightarrow P(A) = 0.240 \\ n = 200 &\Rightarrow P(A) = 0.422 \\ n = 300 &\Rightarrow P(A) = 0.561 \end{aligned}$$

- Therefore, what is your opinion about $P(A)$?

ILLUSTRATIVE EXAMPLE: FREQUENTIST APPROACH

Light bulb lifetime $\Rightarrow X \sim \mathcal{E}(\lambda)$ & $f(x;\lambda) = \lambda e^{-\lambda x} \qquad x, \lambda > 0$

- Sample $\underline{X} = (X_1, \ldots, X_n)$, i.i.d. $\mathcal{E}(\lambda)$
- Likelihood $l_x(\lambda) = \prod_{i=1}^{n} f(X_i;\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$
- MLE: $\widehat{\lambda} = n / \sum_{i=1}^{n} X_i$, C.I., UMVUE, consistency, etc.

What about available prior information on light bulbs behavior?
How can we translate it? $\Rightarrow$ model and **parameter**

ILLUSTRATIVE EXAMPLE: BAYESIAN APPROACH

Light bulb lifetime $\Rightarrow X \sim \mathcal{E}(\lambda)$ & $f(x;\lambda) = \lambda e^{-\lambda x} \qquad x, \lambda > 0$

- Sample $\underline{X} = (X_1, \ldots, X_n)$, i.i.d. $\mathcal{E}(\lambda)$
- Likelihood $l_x(\lambda) = \prod_{i=1}^{n} f(X_i;\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$
- Prior $\lambda \sim \mathcal{G}(\alpha, \beta)$, $\pi(\lambda) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$
- Posterior $\pi(\lambda|\underline{X}) \propto \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i} \cdot \lambda^{\alpha-1} e^{-\beta\lambda}$

  $\Rightarrow \lambda|\underline{X} \sim \mathcal{G}(\alpha + n, \beta + \sum_{i=1}^{n} X_i)$

Posterior distribution fundamental in Bayesian analysis

CONJUGATE PRIORS

- We just saw that a gamma prior on the parameter of an exponential model leads to a gamma posterior
- $\Rightarrow$ The gamma distribution is a conjugate prior for the exponential model
- Does conjugacy occur always? Unfortunately not and simulation methods, e.g. MCMC (Markov chain Monte Carlo), are needed to get samples from the posterior distribution
- There are some relevant cases of conjugacy and we will see some of them:
  - Beta prior conjugate w.r.t. Bernoulli, binomial, geometric models
  - Dirichlet prior conjugate w.r.t. multinomial model
  - Gamma prior conjugate w.r.t. exponential, Poisson models
  - Gaussian prior conjugate w.r.t. Gaussian model with fixed variance/covariance matrix and unknown mean
  - Gaussian-Inverse gamma prior w.r.t. univariate Gaussian model with unknown mean and variance

CONJUGATE PRIOR FOR BINOMIAL

- Binomial data (x "successes" in n trials), with $P(success) = \theta$
  $\Rightarrow l_x(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

- Beta prior $\mathcal{B}e(\alpha,\beta)$: $\pi(\theta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, $0 < \theta < 1$, $\alpha, \beta > 0$

- $\Rightarrow$ posterior $\pi(\theta|x,n) \propto \theta^x(1-\theta)^{n-x} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$

- $\Rightarrow \theta|x,n \sim \mathcal{B}e(\alpha+x, \beta+n-x)$

- Note that the result is proved without using the constant values

- It is worth trying with the following models:
  - Bernoulli: $f(x|\theta) = \theta^x(1-\theta)^{1-x}, x = 0, 1$
  - Geometric: $(1-\theta)\theta^x$, x nonnegative integer

CONJUGATE PRIOR FOR GAUSSIAN

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
- Mean/median $\mu \in \Re$ unknown and variance $\sigma^2 > 0$ known
- $\underline{X} = (X_1, \ldots, X_n)$
- Likelihood:

$$\begin{aligned} L(\underline{X}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/(2\sigma^2)} \end{aligned}$$

- Prior: $\mu \sim \mathcal{N}(\mu_0, \tau^2) \Rightarrow \pi(\mu) = \frac{1}{\sqrt{2\pi}\tau} e^{-(\mu-\mu_0)^2/(2\tau^2)}$

CONJUGATE PRIOR FOR GAUSSIAN

- Posterior:

$$\begin{aligned}
\pi(\mu|\underline{X}) &\propto e^{-\sum_{i=1}^{n}(X_i-\mu)^2/(2\sigma^2)} \cdot e^{-(\mu-\mu_0)^2/(2\tau^2)} \\
&\propto e^{-(n\mu^2-2\mu\sum_{i=1}^{n}X_i)/(2\sigma^2)} \cdot e^{-(\mu^2-2\mu_0\mu)/(2\tau^2)} \\
&\propto e^{-\left\{\mu^2(n/\sigma^2+1/\tau^2)-2\mu(\sum_{i=1}^{n}X_i/\sigma^2+\mu_0/\tau^2)\right\}/2} \\
&\propto \exp\left\{-\frac{1}{2(n/\sigma^2+1/\tau^2)^{-1}}\left[\mu^2-2\mu\frac{\sum_{i=1}^{n}X_i/\sigma^2+\mu_0/\tau^2}{n/\sigma^2+1/\tau^2}\right]\right\}
\end{aligned}$$

$$\Rightarrow \mu|\underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n}X_i/\sigma^2+\mu_0/\tau^2}{n/\sigma^2+1/\tau^2}, \frac{1}{n/\sigma^2+1/\tau^2}\right)$$

- Prior mean: $E(\mu) = \mu_0$
- MLE: $\dfrac{\sum_{i=1}^{n}X_i}{n}$
- Posterior mean: $\dfrac{\sum_{i=1}^{n}X_i/\sigma^2+\mu_0/\tau^2}{n/\sigma^2+1/\tau^2}$

PARAMETER ESTIMATION - DECISION ANALYSIS

- Loss function $L(\lambda, a)$, $a \in \mathcal{A}$ action space
- Minimize $\mathcal{E}^{\pi(\lambda|\underline{X})} L(\lambda, a) = \int L(\lambda, a)\pi(\lambda|\underline{X})d\lambda$ w.r.t. a
  $\Rightarrow \hat{\lambda}$ Bayesian optimal estimator of λ
  - $\hat{\lambda}$ posterior median if $L(\lambda, a) = |\lambda - a|$
  - $\hat{\lambda}$ posterior mean $\mathcal{E}^{\pi(\lambda|\underline{X})}\lambda$ if $L(\lambda, a) = (\lambda - a)^2$

$$\begin{aligned}
\mathcal{E}^{\pi(\lambda|\underline{X})} L(\lambda, a) &= \int (\lambda - a)^2 \pi(\lambda|\underline{X})d\lambda \\
&= \int \lambda^2 \pi(\lambda|\underline{X})d\lambda - 2a \int \lambda\pi(\lambda|\underline{X})d\lambda + a^2 \cdot 1 \\
&= \int \lambda^2 \pi(\lambda|\underline{X})d\lambda - 2a\mathcal{E}^{\pi(\lambda|\underline{X})}\lambda + a^2
\end{aligned}$$

  - $\dfrac{\partial \mathcal{E}^{\pi(\lambda|\underline{X})} L(\lambda, a)}{\partial a} = 0 \Leftrightarrow a = \mathcal{E}^{\pi(\lambda|\underline{X})}\lambda$
  - $\dfrac{\partial^2 \mathcal{E}^{\pi(\lambda|\underline{X})} L(\lambda, a)}{\partial^2 a} = 2 \Rightarrow \mathcal{E}^{\pi(\lambda|\underline{X})}\lambda$ minimum

PRIOR AND DATA INFLUENCE

- Sample $(X_1, \ldots, X_n)$ from $X \sim \mathcal{E}(\lambda)$ with prior $\lambda \sim \mathcal{G}(\alpha, \beta)$
- Posterior mean: $\widehat{\lambda} = \dfrac{\alpha + n}{\beta + \sum X_i}$
- Prior mean: $\widehat{\lambda}_P = \dfrac{\alpha}{\beta}$ (and variance $\sigma^2 = \dfrac{\alpha}{\beta^2}$)
- MLE: $\widehat{\lambda}_M = n / \sum X_i$
- $\alpha_1 = k\alpha$ and $\beta_1 = k\beta \Rightarrow \widehat{\lambda}_{1P} = \widehat{\lambda}_P$ and $\sigma_1^2 = \sigma^2/k$
- Posterior mean: $\widehat{\lambda} = \dfrac{k\alpha + n}{k\beta + \sum X_i}$
- $k \to 0 \Rightarrow$ prior variance $\to \infty \Rightarrow \widehat{\lambda} \to n / \sum X_i$, i.e. MLE (prior does not count)
- $k \to \infty \Rightarrow$ prior variance $\to 0 \Rightarrow \widehat{\lambda} \to \widehat{\lambda}_P$, i.e. prior mean (data do not count)
- $n \to \infty \Rightarrow \widehat{\lambda} \sim \dfrac{n}{\sum X_i}$, i.e. MLE (prior does not count)

PRIOR CHOICE

Where to start from?

- $X \sim \mathcal{E}(\lambda)$
- $f(x|\lambda) = \lambda \exp\{-\lambda x\}$
- $P(X \leq x) = F(x) = 1 - S(x) = 1 - \exp\{-\lambda x\}$

$\Rightarrow$ *Physical* properties of λ

- $\mathbf{E}X = 1/\lambda$
- $VarX = 1/\lambda^2$
- $h(x) = \dfrac{f(x)}{S(x)} = \dfrac{\lambda \exp\{-\lambda x\}}{\exp\{-\lambda x\}} = \lambda$ (hazard function)

PRIOR CHOICE

Possible available information

- Exact prior $\pi(\lambda)$ (???)
- Quantiles of X_i, i.e. $P(X_i \leq x_q) = q$
  - Results from previous experiments (e.g. 75% of light bulbs had failed after 2 years of operation $\Rightarrow$ 2 years is the 75% quantile of X_i)
- Quantiles or moments of λ, i.e. $P(\lambda \leq \lambda_q) = q$ or $\mathbf{E}\lambda^k = a_k$
- Most likely value and upper and lower bounds
- *Expected* value of λ and *confidence* on such value (mean and variance)
  - $\mathcal{E}\lambda = \mu = \frac{\alpha}{\beta}$ and $Var\lambda = \sigma^2 = \frac{\alpha}{\beta^2} \Rightarrow \alpha = \frac{\mu^2}{\sigma^2}$ and $\beta = \frac{\mu}{\sigma^2}$
- None of them

PRIOR CHOICE

Which prior?

- $\lambda \sim \mathcal{G}(\alpha, \beta) \Rightarrow f(\lambda|\alpha, \beta) = \beta^{\alpha}\lambda^{\alpha-1}\exp\{-\beta\lambda\}/\Gamma(\alpha)$ (conjugate)
- $\lambda \sim \mathcal{LN}(\mu, \sigma^2) \Rightarrow f(\lambda|\mu, \sigma^2) = \{\lambda\sigma\sqrt{2\Pi}\}^{-1}\exp\{-(\log\lambda - \mu)^2/(2\sigma^2)\}$
- $\lambda \sim \mathcal{GEV}(\mu, \sigma, \theta) \Rightarrow f(\lambda) = \frac{1}{\sigma}\left[1 + \theta\left(\frac{\lambda-\mu}{\sigma}\right)\right]_+^{-1/\theta-1} \exp\left\{-\left[1 + \theta\left(\frac{\lambda-\mu}{\sigma}\right)\right]_+^{-1/\theta}\right\}$
- $\lambda \sim \mathcal{T}(l, m, u)$ (triangular)
- $\lambda \sim \mathcal{U}(l, u)$
- $\lambda \sim \mathcal{W}(\mu, \alpha, \beta) \Rightarrow f(\lambda) = \frac{\beta}{\alpha}\left(\frac{\lambda-\mu}{\alpha}\right)^{\beta-1}\exp\{-\left(\frac{\lambda-\mu}{\alpha}\right)^{\beta}\}$
- ...

CREDIBLE INTERVALS

- In Bayesian statistics the parameter λ is considered a r.v. and it is possible to compute the posterior probability $\mathcal{P}(\lambda \in A|\underline{X})$ for a measurable set A

- $\Rightarrow$ Credible set, as a counterpart of the frequentist confidence set, but with very different meaning

- If the set is an interval, then we call it *credible interval at* $100y\%$, if its posterior probability is y

- We are interested also in the *highest posterior density (HPD) sets*, which are the ones with the smallest Lebesgue measure among those with a given posterior probability

- Light bulb: $\mathcal{P}(\lambda \leq z|\underline{X}) = \int_0^z \frac{(\beta + \sum X_i)^{\alpha+n}}{\Gamma(\alpha+n)} \lambda^{\alpha+n-1} e^{-(\beta+\sum X_i)\lambda} d\lambda$

CREDIBLE INTERVALS

- One observation $X \sim \mathcal{N}(\mu, 1)$
- Prior $\mu \sim \mathcal{N}(0, 1)$
- Posterior
  $\pi(\mu|x) \propto e^{-(x-\mu)^2/2} \cdot e^{-\mu^2/2} \propto e^{-(\mu^2 - x\mu)} \propto \exp\{\frac{1}{2 \cdot 1/2}(\mu - x/2)^2\}$
  $\Rightarrow \mu|x \sim \mathcal{N}(x/2, 1/2)$
- $Z = \dfrac{\mu - x/2}{\sqrt{1/2}} \sim \mathcal{N}(0, 1)$
- Quantiles $Z_{.975} = 1.96$ and $Z_{.025} = -1.96$
- $\Rightarrow P\left(Z_{.025} \leq Z \leq Z_{.975}\right) = \left(-1.96 \leq \dfrac{\mu - x/2}{\sqrt{1/2}} \leq 1.96\right) = .95$
- $\Rightarrow \left(x/2 - 1.96\sqrt{1/2}, x/2 + 1.96\sqrt{1/2}\right)$ credible interval at 95% for μ

HYPOTHESIS TESTING

- $H_0 : \lambda \in \Lambda_0$ vs. $H_1 : \lambda \in \Lambda_0^C$, where C denotes the complement set
- Priors: $\mathbb{P}(H_0) = \mathbb{P}(\lambda \in \Lambda_0) = 1 - \mathbb{P}(\lambda \in \Lambda_0^C) = 1 - \mathbb{P}(H_1)$
- Sample $\underline{X} \Rightarrow$ posteriors $\mathbb{P}(H_0|\underline{X}) = 1 - \mathbb{P}(H_1|\underline{X})$
- There are many problems associated with the frequentist approach to hypothesis testing which can be addressed properly in a Bayesian framework
  - Bayesians have no need to know if either H_0 or H_1 is true but, treating λ as a r.v., they can assess the probabilities of both hypotheses and decide based on them
  - Frequentists are unable to specify opinions about hypotheses, unlike Bayesians with prior distributions on them
  - Frequentists set significance levels a priori and decide based on them, unlike Bayesians which get a posteriori the probability of an hypothesis and decide based on it

HYPOTHESIS TESTING

- One sided test: $H_0 : \lambda \leq \lambda_0$ vs. $H_1 : \lambda > \lambda_0$

  $\Rightarrow$ Reject H_0 iff $\mathbb{P}(\lambda \leq \lambda_0 | \underline{X}) \leq \alpha$, α significance level (e.g. $\alpha = 0.5$)

- Two sided test: $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$
  - Problems with $\mathbb{P}(\lambda = \lambda_0 | \underline{X})$
  - Do not reject if $\lambda_0 \in A$, A $100(1-\alpha)\%$ credible interval
  - Consider $\mathbb{P}([\lambda_0 - \epsilon, \lambda_0 + \epsilon] | \underline{X})$
  - Dirac measure: $\mathbb{P}(\lambda_0) > 0$ and consider $\mathbb{P}(\lambda_0 | \underline{X})$

PREDICTION

- After observing an i.i.d. sample $\underline{X} = (X_1, \ldots, X_n)$, what can we say about a next observation X_{n+1} from the same density $f(X|\lambda)$?

- We could consider the next observations $X_{n+1}, \ldots, X_{n+j}$ but we take $j = 1$ for simplicity

- When considering observations over time we prefer to use the term *forecast* instead of *prediction* (e.g., weather forecast)

- Given the sample $\underline{X}$ and the prior $\pi(\lambda)$, then the posterior $\pi(\lambda|\underline{X})$ is used to compute the posterior predictive density (absolutely continuous case here) for X_{n+1}
  $f(X_{n+1}|\underline{X}) = \int f(X_{n+1}|\lambda, \underline{X})\pi(\lambda|\underline{X})d\lambda = \int f(X_{n+1}|\lambda)\pi(\lambda|\underline{X})d\lambda$

- Prior predictive densities can be used to compare model via Bayes factor (more later)

- Posterior predictive densities can be used to assess the goodness of fit of a model through the prediction error, using part of the data to get the posterior and the remaining one to get predicted values (e.g. predicted posterior mean/median) and compare them with actual ones

PREDICTION

- Light bulb: $X_{n+1}|\lambda \sim \mathcal{E}(\lambda),\ \lambda|\underline{X} \sim \mathcal{G}(\alpha+n, \beta+\sum X_i)$

- Posterior predictive density for X_{n+1}

$$
\begin{aligned}
f_{X_{n+1}}(X_{n+1}|\underline{X}) &= \int_0^\infty \lambda e^{-\lambda X_{n+1}} \cdot \frac{(\beta+\sum X_i)^{\alpha+n}}{\Gamma(\alpha+n)} \lambda^{\alpha+n-1} e^{-\lambda(\beta+\sum X_i)} d\lambda \\
&= \frac{(\beta+\sum X_i)^{\alpha+n}}{\Gamma(\alpha+n)} \int_0^\infty \lambda^{\alpha+n+1-1} e^{-\lambda(\beta+\sum X_i+X_{n+1})} d\lambda \\
&= \frac{(\beta+\sum X_i)^{\alpha+n}}{\Gamma(\alpha+n)} \frac{\Gamma(\alpha+n+1)}{(\beta+\sum X_i+X_{n+1})^{\alpha+n+1}} \\
&= (\alpha+n) \frac{(\beta+\sum X_i)^{\alpha+n}}{(\beta+\sum X_i+X_{n+1})^{\alpha+n+1}}
\end{aligned}
$$

- I found first the constant knowing that the density integrates to 1 and then I used the property $\Gamma(z+1) = z\Gamma(z)$

MODEL SELECTION

Compare $\mathcal{M}_1 = \{f_1(x|\theta_1), \pi(\theta_1)\}$ and $\mathcal{M}_2 = \{f_2(x|\theta_2), \pi(\theta_2)\}$

- Bayes factor

$$\Rightarrow BF = \frac{f_1(x)}{f_2(x)} = \frac{\int f_1(x|\theta_1)\pi(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi(\theta_2)d\theta_2}$$

BF	$2\log_{10} BF$	Evidence in favor of $\mathcal{M}_1$
1 to 3	0 to 2	Hardly worth commenting
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

- Posterior odds

$$\Rightarrow \frac{P(\mathcal{M}_1|data)}{P(\mathcal{M}_2|data)} = \frac{P(data|\mathcal{M}_1)}{P(data|\mathcal{M}_2)} \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} = BF \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}$$

BAYESIAN ROBUSTNESS: MOTIVATING EXAMPLE

- $X \sim \mathcal{N}(\theta, 1)$
- Expert's opinion on prior P: median at 0, quartiles at ± 1, symmetric and unimodal
- $\Rightarrow$ Possible priors include Cauchy $\mathcal{C}(0, 1)$ and Gaussian $\mathcal{N}(0, 2.19)$
- Interest in posterior mean $\mu^C(x)$ or $\mu^N(x)$

x	0	1	2	4.5	10
$\mu^C(x)$	0	0.52	1.27	4.09	9.80
$\mu^N(x)$	0	0.69	1.37	3.09	6.87

- Decision strongly dependent on the choice of the prior for large x
- Robust alternative: Posterior median w.r.t. posterior mean
- Range of posterior mean in class of priors compatible with expert's opinions

MARKOV CHAIN MONTE CARLO (MCMC)

- It is not always possible to get posterior distributions in closed form
- Use of Bayesian Statistics limited until early 90's "rediscovery" of MCMC
- The name MCMC is due to the Monte Carlo simulation applied to a Markov chain whose stationary distribution is, under adequate conditions, the posterior distribution
- *Gibbs sampling*, the simplest MCMC, with simulation based on full posterior conditional distributions, for each parameter given the others and the data
- Suppose the parameter is $\theta = (\mu, \tau)$ and the data are $\underline{X}$, then the simulation is based on N replications of the following steps, with $i = 0$, $\mu^{(0)} = \mu_0$ and $\tau^{(0)} = \tau_0$
  1. $\mu^{(i+1)} \sim \pi(\mu|\tau^{(i)}, \underline{X})$
  2. $\tau^{(i+1)} \sim \pi(\tau|\mu^{(i+1)}, \underline{X})$
  3. $i = i + 1$; if $i \leq N$ then go to 1
- The posterior distributions approximated by histograms of μ's and τ's
- Posterior quantities can be computed, e.g. $E(\mu|\underline{X}) = (1/N) \sum_{i=1}^{N} \mu^{(i)}$

MARKOV CHAIN MONTE CARLO (MCMC)

- Sample $\underline{X}$ and parameter $\theta = (\theta_1, \ldots, \theta_n)$
- Notation $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, for $i = 1, \ldots, n$
- *Gibbs sampling* used if $\pi(\theta|\underline{X})$ unavailable but all $\pi(\theta_i|\theta_{-i}, \underline{X})$, $i = 1, \ldots, n$, are
- Consider $X_i \sim \mathcal{N}(\mu, \tau)$, $i = 1, \ldots, n$
- $\tau = 1/\sigma^2$ precision, $\underline{X} = (X_1, \ldots, X_N)$, $\overline{X} = \sum_{i=1}^N X_i/N$ sample mean
- Likelihood: $\prod_{i=1}^N \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(X_i-\mu)^2}$
- Prior distribution: $\pi(\mu, \tau) \propto \tau^{a-1} e^{-b\tau}$: is it "strange"?
- Posterior: $\propto \tau^{a+N/2-1} e^{-\tau(b+\sum_{i=1}^N (X_i-\mu)^2/2)}$
- $\mu|\tau, \underline{X} \sim \mathcal{N}(\overline{X}, n\tau)$ and $\tau|\mu, \underline{X} \sim \mathcal{G}(a + N/2, b + \sum_{i=1}^N (X_i - \mu)^2/2)$

MARKOV CHAIN MONTE CARLO

- In words, Gibbs sampling consists of a "sufficient" number of steps in which each parameter θ_i is sequentially drawn from its *full conditional distribution* $\pi(\theta_i|\theta_{-i}, \underline{X})$, where θ_{-i} contains the values of $\theta_1, \ldots, \theta_{i-1}$ generated at the current step and those of $\theta_{i+1}, \ldots, \theta_n$ generated at the previous step

- Algorithm

  1. Set $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$ and $j = 0$
  2. Set $j = j + 1$
  3. For $i = 1, \ldots, n$, draw $\theta_i^{(j)}$ from $\pi(\theta_i|\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_n^{(j-1)}, \underline{X})$
  4. If $j < N$ (set a priori) then go back to (2)
  5. $\Rightarrow \theta^{(j)}, j = 1, \ldots, N$, used to get a sample from the posterior distribution

- Some $\theta^{(j)}$'s might be discarded, e.g. initial ones (more later)

MARKOV CHAIN MONTE CARLO *

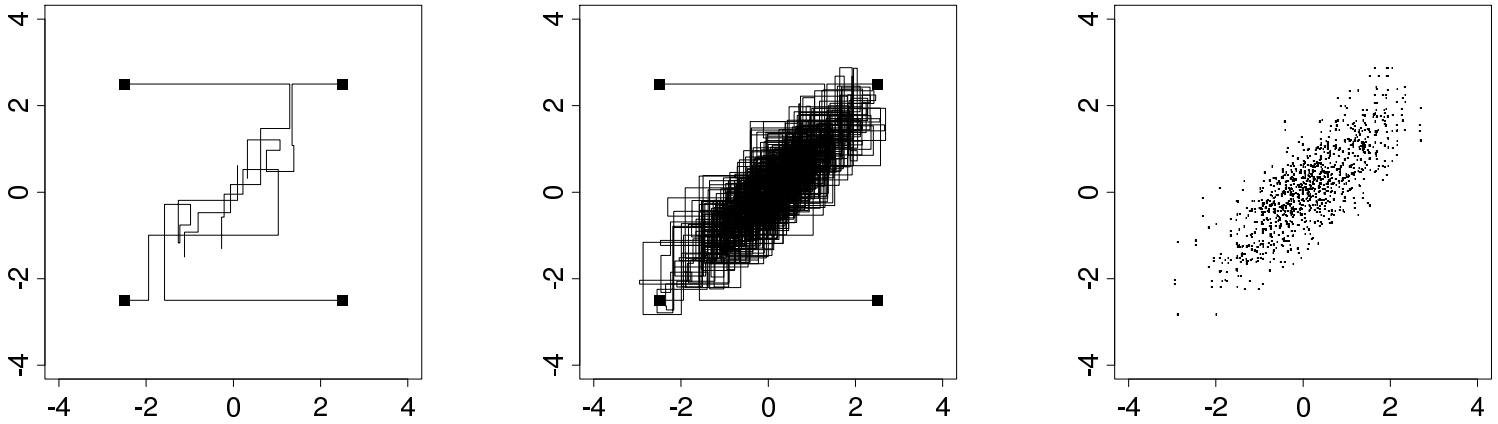
- Consider a single observation (y_1, y_2) from a bivariate Gaussian with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$
- Uniform prior on θ: $\pi(\theta) \propto c,\ c > 0$
- $\Rightarrow$ Posterior $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} | y \sim \mathcal{N}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$
- Although it is simple to draw directly from the joint posterior distribution of (θ_1, θ_2), for the purpose of exposition we demonstrate the Gibbs sampler here
- Simulate (alternating) from known full conditional distributions
  - $\theta_1 | \theta_2, y \sim \mathcal{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$
  - $\theta_2 | \theta_1, y \sim \mathcal{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$

*Example from Gelman et al., *Bayesian Data Analysis, Third Edition*, freely available at `http://www.stat.columbia.edu/~gelman/book/BDA3.pdf`

MARKOV CHAIN MONTE CARLO

- Take $\rho = 0.8$ and $(y_1, y_2) = (0, 0)$
- $\Rightarrow$ Posterior $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} | y \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right)$
- Four independent sequences starting at $(\pm 2.5, \pm 2.5)$ to remove dependence on initial point
- Sequences run until convergence to the posterior is achieved (more later on checking for convergence)
- By convergence we mean that the drawn samples are from an approximating distribution close to the posterior one (our target)
- Use of just part of the data, removing the initial ones since they might not be in the approximating distribution (this operation is called *burn-in*)
- Sometimes one aims to reduce correlation between samples so that just 1 every m is kept

MARKOV CHAIN MONTE CARLO



- Left: First 10 iterations for four independent sequences starting at $(\pm 2.5, \pm 2.5)$
- Center: After 500 iterations, the sequences have reached approximate convergence
- Right: The points from the second halves of the sequences, discarding the first 250 samples values of each sequence (burn-in)
- Often just one sequence is drawn but for longer time
- Note how the samples are around $(0, 0)$ and showing a strong positive correlation, as expected knowing the exact joint posterior

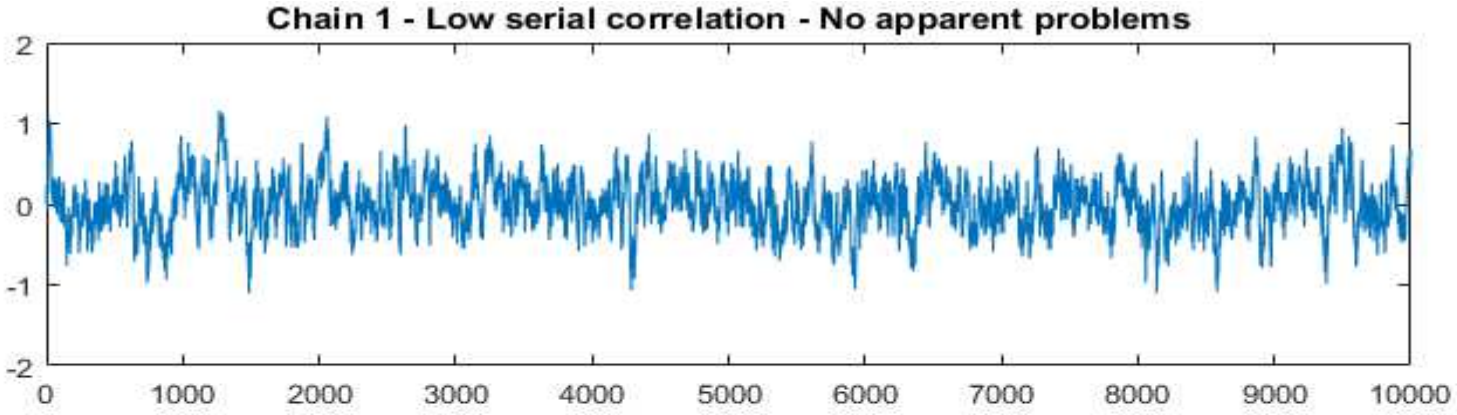
MARKOV CHAIN MONTE CARLO

- In Gibbs sampling we assumed that it was always possible to get the full conditional $\pi(\theta_i|\theta_{-i}, \underline{X})$ for all i's but is not always the case
- Sometimes we know only $\pi(\theta_i|\theta_{-i}, \underline{X}) \propto q(\theta_i|\theta_{-i}, \underline{X})$ where $q(\cdot)$ is not a density function
- In this case we will use *Metropolis-Hastings steps within Gibbs*
- The Metropolis-Hastings algorithm allows to draw a value θ_i^* from a proposal density $p(\theta_i)$ and accept either it or $\theta_i^{(j-1)}$ as $\theta_i^{(j)}$ with probabilities depending on both p and q
- The proposal density for θ_i^* could be chosen, e.g., either as the same for each iteration or as dependent on the previous $\theta_i^{(j-1)}$

MARKOV CHAIN MONTE CARLO

- We already saw that running more than one simulation at the time and removing the initial values should reduce the dependence on the initial values
- The proposal distributions are often chosen depending on the value at the previous iteration, e.g. a Gaussian distribution centered at it, or independently from it, possibly the same at all iterations, e.g. Gaussians with the same mean
- Many tools developed to check convergence of the sequence to the true distribution
- The simplest, graphical, tool to assess convergence is to check if the plot of the sample mean stabilises as the iterations grow (if not, then no convergence)
- Given a sample $\theta^{(S+1)}, \ldots, \theta^{(N)}$, with a burn-in of size S, then estimators of $E(h(\theta)|y)$ are given by $\frac{\sum_{j=S+1}^{N} h(\theta^{(j)})}{N-S}$, like
  - $E(\theta|y) \approx \frac{\sum_{j=S+1}^{N} \theta^{(j)}}{N-S}$
  - $\mathbb{P}(\theta \in A|y) \approx \frac{\#\{\theta^{(j)} \in A\}_{j=S+1}^{N}}{N-S}$

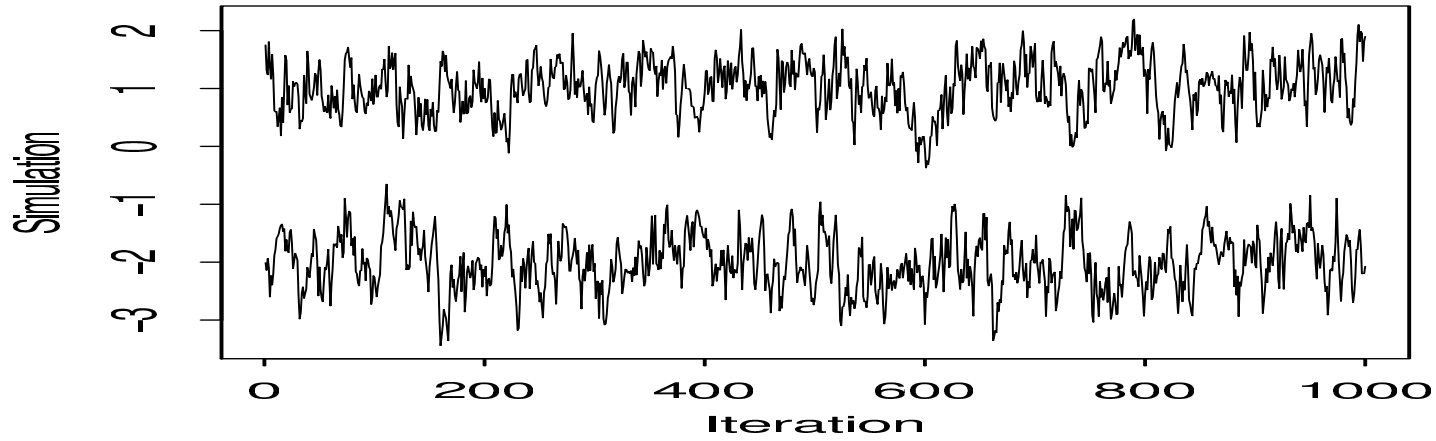
MARKOV CHAIN MONTE CARLO*



- Trace plots are heuristic tools, widely used to check convergence of the MCMC
- They plot the values of each parameter for all the iterations
- They are ”good” when the plot keeps jumping within a set which denotes where the posterior density is concentrated
- The trace plot in the figure is a good one, unlike the next ones

*Plots from `www.statlect.com`

MARKOV CHAIN MONTE CARLO

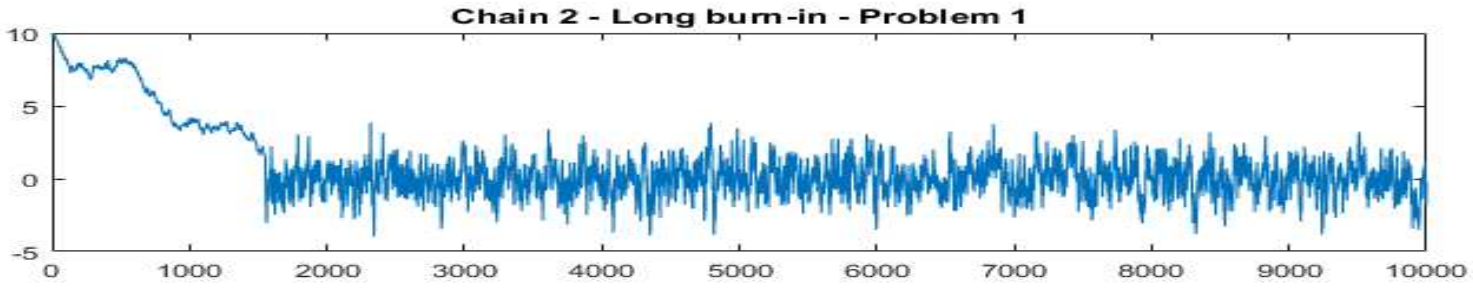


- Here two sequences have been running and both of them are converging but to two different values

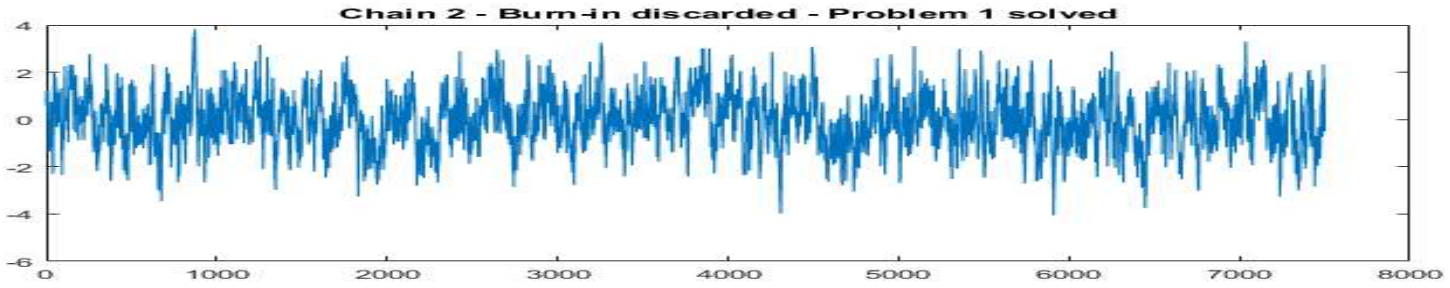
- In general, a plot like this is not desirable since it does not give a clear indication about where the posterior density is, unless the density is bimodal

- In the latter case one would expect the chain to jump from one mode to another

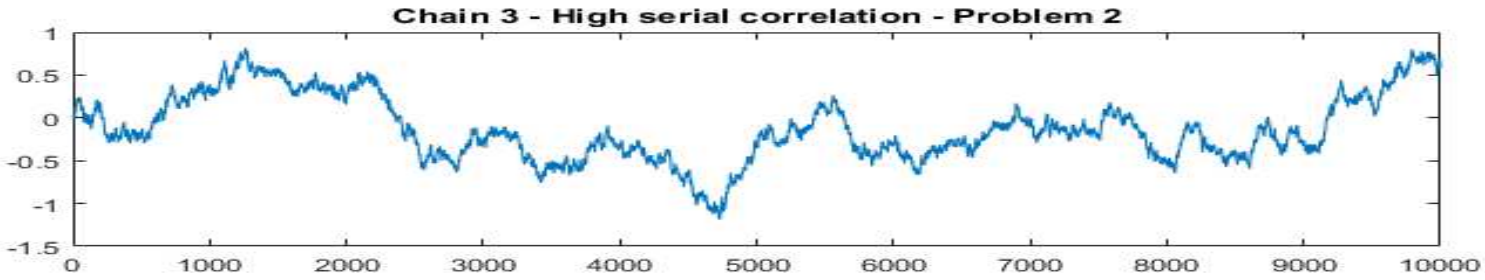
MARKOV CHAIN MONTE CARLO



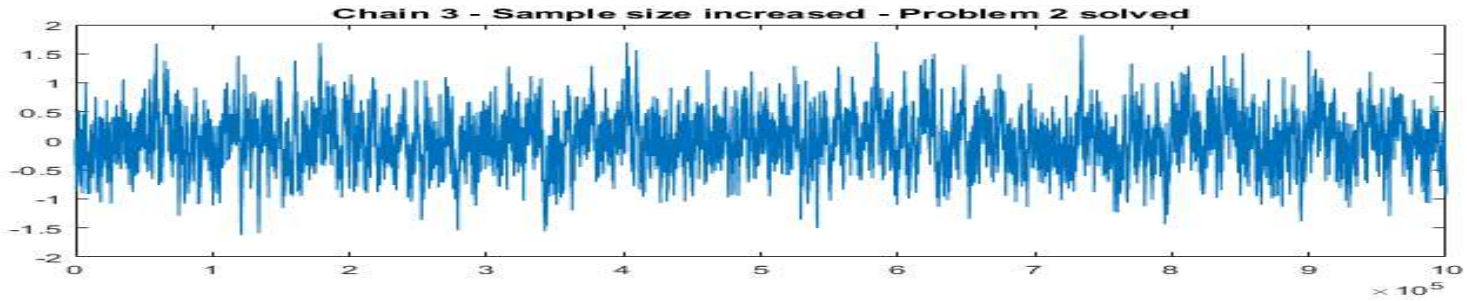
- The first part of the sample looks very different from the remaining part.
- Most likely, the initial distribution and the distributions of the subsequent terms of the chain were very different from the target distribution, but then the chain slowly converged to the target distribution
- The problem can be solved by removing the initial values (burn-in)



MARKOV CHAIN MONTE CARLO



- A lot of autocorrelation between the draws ($\Rightarrow$ lack of independence)
- Chain very slow in exploring the sample space, explored only few times
- The problem could be due to a small number of iterations $\Rightarrow$ run longer and, possibly, take one draw out of m to avoid large sample size and remove autocorrrelation



REGRESSION

- We now consider linear regression (LR), providing a linear relation between a dependent variable (Y) and an independent one (X), sometimes called *covariate*
- We can distinguish 4 cases based on the dimensions of Y and X
  - Simple LR vs. Multiple LR: just one X or multiple X's
  - Univariate LR vs. Multivariate LR: just one-dimensional Y or multiple dimensional Y
- We consider only the simplest case: Univariate Simple Linear Regression
- $Y = \beta_1 + \beta_2 X + \varepsilon$
- β_1, β_2 univariate unknown parameters
- ε error term with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$ unknown
- We consider $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

REGRESSION

- Observations: $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, $i = 1, \ldots, n$
- X_i's are supposed known here but they could be r.v.'s as well
- We assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$
- Notation: $\underline{Y} = (Y_1, \ldots, Y_n)$ and $\underline{X} = (X_1, \ldots, X_n)$
- Likelihood function $L(\beta_1, \beta_2, \sigma^2 | \underline{Y}, \underline{X})$ given by

$$\begin{aligned} \prod_{i=1}^{n} f(Y_i | X_i, \beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2}\} \right\} \\ &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\{-\frac{\sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2}\} \end{aligned}$$

- Independent priors with known hyperparameters:
  $\beta_1 \sim \mathcal{N}(0, \tau_1^2)$, $\beta_2 \sim \mathcal{N}(0, \tau_2^2)$ and $\sigma^2 \sim \mathcal{IG}(a, b)$

REGRESSION

- Posterior distribution

$$\begin{aligned}\pi(\beta_1,\beta_2,\sigma^2|\underline{Y},\underline{X}) &\propto \frac{1}{(\sigma^2)^{n/2}}\exp\{\frac{-\sum_{i=1}^n(Y_i-\beta_1-\beta_2X_i)^2}{2\sigma^2}\}\cdot \\ &\cdot \exp\{-\beta_1^2/(2\tau_1^2)\}\exp\{-\beta_2^2/(2\tau_2^2)\}\frac{1}{(\sigma^2)^{a+1}}\exp\{-b/\sigma^2\}\end{aligned}$$

- Conditional on β_1: $\beta_1|\beta_2,\sigma^2,\underline{Y},\underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n(Y_i-\beta_2X_i)}{n+\sigma^2/\tau_1^2}, \frac{1}{n/\sigma^2+1/\tau_1^2}\right)$

$$\begin{aligned}\pi(\beta_1|\beta_2,\sigma^2,\underline{Y},\underline{X}) &\propto \exp\{\frac{-(n\beta_1^2-2\beta_1\sum_{i=1}^n(Y_i-\beta_2X_i)}{2\sigma^2}\}\exp\{-\beta_1^2/(2\tau_1^2)\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\tau_1^2}\right)\beta_1^2-2\frac{\beta_1}{\sigma^2}\sum_{i=1}^n(Y_i-\beta_2X_i)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2(n/\sigma^2+1/\tau_1^2)^{-1}}\left[\beta_1^2-2\frac{\beta_1}{\sigma^2}\frac{\sum_{i=1}^n(Y_i-\beta_2X_i)}{n/\sigma^2+1/\tau_1^2}\right]\right\}\end{aligned}$$

REGRESSION

- Conditional on β_2: $\beta_2|\beta_1,\sigma^2,\underline{Y},\underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n X_i(Y_i-\beta_1)}{\sum_{i=1}^n X_i^2+\sigma^2/\tau_1^2}, \frac{1}{\sum_{i=1}^n X_i^2/\sigma^2+1/\tau_1^2}\right)$

$$\begin{aligned}
\pi(\beta_2|\beta_1,\sigma^2,\underline{Y},\underline{X}) &\propto \exp\{-\frac{\beta_2^2\sum_{i=1}^n X_i^2-2\beta_2\sum_{i=1}^n X_i(Y_i-\beta_1)}{2\sigma^2}\}\exp\{-\beta_2^2/(2\tau_2^2)\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{\sum_{i=1}^n X_i^2}{\sigma^2}+\frac{1}{\tau_2^2}\right)\beta_2^2-2\frac{\beta_2}{\sigma^2}\sum_{i=1}^n X_i\left(Y_i-\beta_1\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2(\sum_{i=1}^n \frac{X_i^2}{\sigma^2}+\frac{1}{\tau_1^2})^{-1}}\left[\beta_2^2-2\frac{\beta_2}{\sigma^2}\frac{\sum_{i=1}^n X_i\left(Y_i-\beta_1\right)}{\sum_{i=1}^n X_i^2/\sigma^2+1/\tau_1^2}\right]\right\}
\end{aligned}$$

- Conditional on σ^2: $\sigma^2|\beta_1,\beta_2,\underline{Y},\underline{X} \sim \mathcal{IG}\left(a+n/2, b+\frac{\sum_{i=1}^n(Y_i-\beta_1-\beta_2X_i)^2}{2}\right)$

$$\pi(\sigma^2|\beta_1,\beta_2,\underline{Y},\underline{X}) \propto \frac{1}{(\sigma^2)^{n/2}}\exp\{-\frac{\sum_{i=1}^n(Y_i-\beta_1-\beta_2X_i)^2}{2\sigma^2}\}\frac{1}{(\sigma^2)^{a+1}}\exp\{-b/\sigma^2\}$$

- $\Rightarrow$ Gibbs sampling

REGRESSION

- We have found the posterior distributions of the parameters in a suitable form to apply MCMC $\Rightarrow$ now we can estimate them, e.g., considering the posterior mean, and build credible intervals in a way similar to what we saw earlier (and I will not repeat it)
- When considering more than one covariate, i.e., $X_1, \ldots, X_n$, still Gaussian priors should be considered for each of them
- Similarly to the frequentist approach, there is an interest for the covariates which are significant
  - Instead of considering p-values, Bayesians look for a credible interval and check if 0 belongs to it
  - If the credible interval contains 0 then the covariate is not significant; otherwise, it is
  - We will see an example next
- If Y is multivariate, then multivariate Gaussian distributions are chosen to model the observations and as a prior for the mean, while an Inverse Wishart distribution is chosen for the covariance matrix

REGRESSION

- Both frequentist and Bayesian methods will be applied in the next example
- 713 observations corresponding to the days where the prices of the Bitcoins in 8 different exchange markets were recorded together with the prices of the classical assets and the exchange rates
- We will use the package `rstanarm` and the function `stan_glm`, whose usage is similar to `lm`
- Use of improper priors leading to results close to frequentist ones
- You could try other priors, using the `R` tutorials, like `?stan_glm`
- For this example, I tried `stan_lm`, the very equivalent of `lm` (both about linear models) but it did not work, so that I used the one for generalised linear models
- I first present the commands for the frequentist analysis

REGRESSION

```
rm(list=ls()) # Clear the environment
install.packages("ggplot2",dependencies=TRUE)
install.packages("readxl",dependencies=TRUE)
install.packages("corrplot",dependencies=TRUE)
library(ggplot2);library(readxl);library(corrplot)
exchanges<-read_excel("exchanges.xlsx") # Read in working directory
data<-exchanges
data1<-data[-1] # Remove the first column from data
# New dataset with returns instead of prices: (log(x)-log(x-1))
data2<-as.data.frame(sapply(data1,function(x)diff(log(x),lag=1)))
attach(data2) # Bring the names of the variables directly into memory
```

REGRESSION

```
# Multiple linear regression [btc_coinbase on all other variables]
model_3<-lm(btc_coinbase~.,data=data2)
summary(model_3)
# Get and plot residuals
res<-model_3$residuals
plot(res,type='l')
install.packages("rstanarm",dependencies=TRUE)
library(rstanarm)
model_b<-stan_glm(btc_coinbase~.,data=data2)
summary(model_b, digits=3)
# Get and plot residuals
resb<-model_b$residuals
plot(resb,type='l')
```

Default priors: standard Gaussian for intercept and coefficients and exponential of parameter 1 for σ^2

REGRESSION

Results based on MLE

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0001059  0.0003123   0.339  0.73454
btc_kraken    0.0210321  0.0123977   1.696  0.09025 .
btc_bitstamp  0.0384385  0.0359272   1.070  0.28503
btc_itbit     0.0130343  0.0256007   0.509  0.61082
btc_bitfinex  0.2297741  0.0315236   7.289 8.47e-13 ***
btc_hitbtc    0.0821093  0.0184755   4.444 1.03e-05 ***
btc_gemini    0.5981632  0.0308680  19.378  < 2e-16 ***
btc_bittrex   0.0056419  0.0145595   0.388  0.69850
usdyuan      -0.1045943  0.2066436  -0.506  0.61291
usdeur        0.2060414  0.0986501   2.089  0.03710 *
gold          0.0712161  0.0575053   1.238  0.21597
oil          -0.0595675  0.0192726  -3.091  0.00208 **
sp500        -0.0952889  0.0569865  -1.672  0.09495 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

REGRESSION

Results based on Bayes

Estimates:

	mean	sd	10%	50%	90%
(Intercept)	0.000	0.000	0.000	0.000	0.001
btc_kraken	0.021	0.012	0.005	0.021	0.037
btc_bitstamp	0.040	0.036	-0.006	0.040	0.086
btc_itbit	0.012	0.026	-0.020	0.012	0.044
btc_bitfinex	0.228	0.031	0.189	0.228	0.268
btc_hitbtc	0.082	0.019	0.057	0.081	0.106
btc_gemini	0.599	0.031	0.560	0.599	0.638
btc_bittrex	0.006	0.015	-0.013	0.006	0.025
usdyuan	-0.103	0.206	-0.360	-0.106	0.162
usdeur	0.205	0.099	0.078	0.205	0.333
gold	0.071	0.058	-0.003	0.070	0.145
oil	-0.059	0.019	-0.084	-0.060	-0.035
sp500	-0.095	0.058	-0.170	-0.095	-0.021
sigma	0.008	0.000	0.008	0.008	0.009

REGRESSION

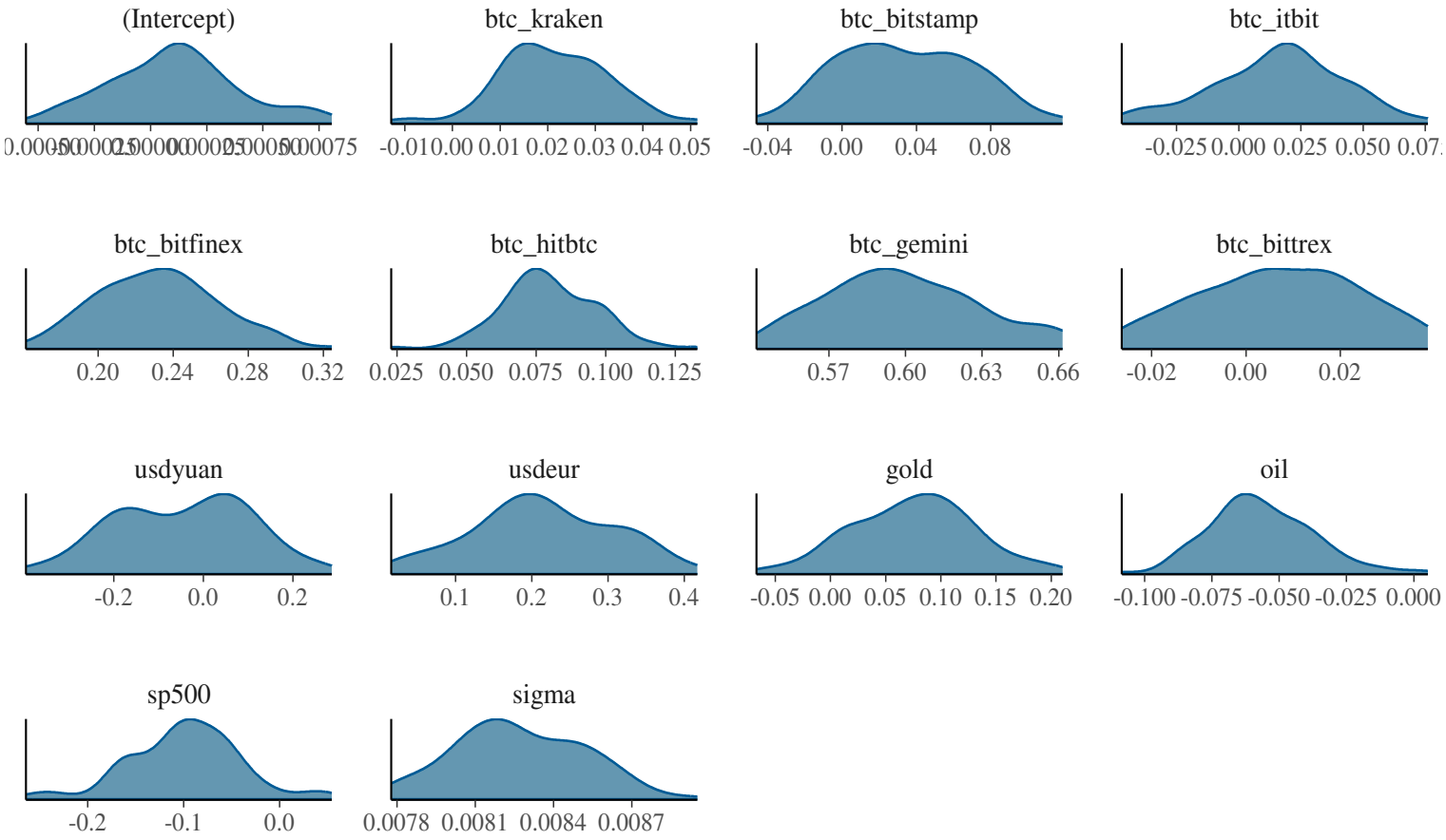
- `Warmup` is better known as *burn-in*, i.e. the first values are discarded because affected by the starting values

- We now consider different priors, like Student t for each coefficient, Cauchy for the intercept and exponential for σ^2

- We consider also 1 chains, setting a seed and the number of iterations

```
model_b<-stan_glm(btc_coinbase~.,chains=1,seed=12345,iter=250,
prior=student_t(df=4,0,2.5),prior_intercept=cauchy(0,10),prior_aux =
exponential(1/2),data=data2)
summary(model_b, digits=3)
print(model_b)
prior_summary(model_b) # To see the chosen priors
library(bayesplot)
mcmc_dens(model_b)
library(bayestestR)
hdi(model_b)
```

REGRESSION



Posterior distributions of all parameters

REGRESSION

Highest density intervals

Parameter	95% HDI
(Intercept)	[0.00, 0.00]
btc_kraken	[0.00, 0.04]
btc_bitstamp	[-0.03, 0.10]
btc_itbit	[-0.04, 0.06]
btc_bitfinex	[0.17, 0.29]
btc_hitbtc	[0.04, 0.11]
btc_gemini	[0.55, 0.66]
btc_bittrex	[-0.02, 0.04]
usdyuan	[-0.28, 0.29]
usdeur	[0.02, 0.36]
gold	[-0.04, 0.18]
oil	[-0.09, -0.01]
sp500	[-0.19, 0.05]

LOGISTIC REGRESSION

- The previous example dealt with continuous variables but what about a response (*dependent variable*) taking only a finite number of integer values?
- Consider people applying for mortgages (or subject to surgery): are they able to pay the mortgage back (or will they survive)?
- The observations are 1's (pays back/survives) and 0's (does not pay back/dies)
- We are still interested in studying the effect of covariates (*independent variables*), like age and gender, on the final result
- We cannot use $Y = \beta_1 + \beta_2 X + \epsilon$ with $Y = 0, 1$ since it is almost impossible to choose r.h.s. terms such that there is always either 0 or 1 in the l.h.s.
- We consider $\pi = P(Y = 1)$ but we cannot use $\pi = \beta_1 + \beta_2 X + \epsilon$ since it is almost impossible to choose r.h.s. terms such that the l.h.s. will be always between 0 and 1
- (`logit`) transformation: $\log\left(\frac{\pi}{1-\pi}\right) = X'\beta$, with X', β vectors of size k
- Earlier: $X' = (1, X), \beta' = (\beta_1, \beta_2)$

LOGISTIC REGRESSION

- $\log\left(\frac{\pi}{1-\pi}\right) = X^{'}\beta \Rightarrow \pi = \frac{e^{X'\beta}}{1+e^{X'\beta}}$
- For each $i = 1, \ldots, n$, consider n_i observations (y_i, x_i) and the related probability π_i (e.g. y_i, out of n_i, persons with features x_i, paid the mortgage back)
- $\underline{y} = (y_1, \ldots, y_n)$, $\underline{x} = (x_1, \ldots, x_n)$, $\underline{\pi} = (\pi_1, \ldots, \pi_n)$ and $\underline{n} = (n_1, \ldots, n_n)$
- We consider a Binomial model (Bernoulli if $n_i = 1$)

$$P(Y_i = y_i|\pi_i, n_i, x_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} = \binom{n_i}{y_i}\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)^{y_i}\left(\frac{1}{1+e^{x_i'\beta}}\right)^{n_i-y_i}$$

- Likelihood: $\prod_{i=1}^{n}\binom{n_i}{y_i}\frac{e^{y_i x_i'\beta}}{\left(1+e^{x_i'\beta}\right)^{n_i}}$
- Prior distribution on β: e.g. Multivariate Gaussian (simplest: product of independent univariate Gaussian distributions)

LOGISTIC REGRESSION

- Survey of 3200 residents in a small area of Bangladesh suffering from arsenic contamination of groundwater*
- Respondents with elevated arsenic levels in their wells were encouraged to switch their water source to a safe well in the nearby area and the survey was conducted several years later to learn which of the affected residents had switched wells
- The goal of the analysis is to learn about the factors associated with switching wells
- To start, we will use `dist` (the distance from the respondent's house to the nearest well with safe drinking water) as the only predictor of `switch` (1 if switched, 0 if not).
- Then we will expand the model by adding the arsenic level of the water in the resident's own well as a predictor and then we will add all variables
- After loading the `wells` data, we first rescale the `dist` variable (measured in meters) so that it is measured in units of 100 meters

*Example due to Gabry and Goodrich (website), based on Gelman and Hill's book

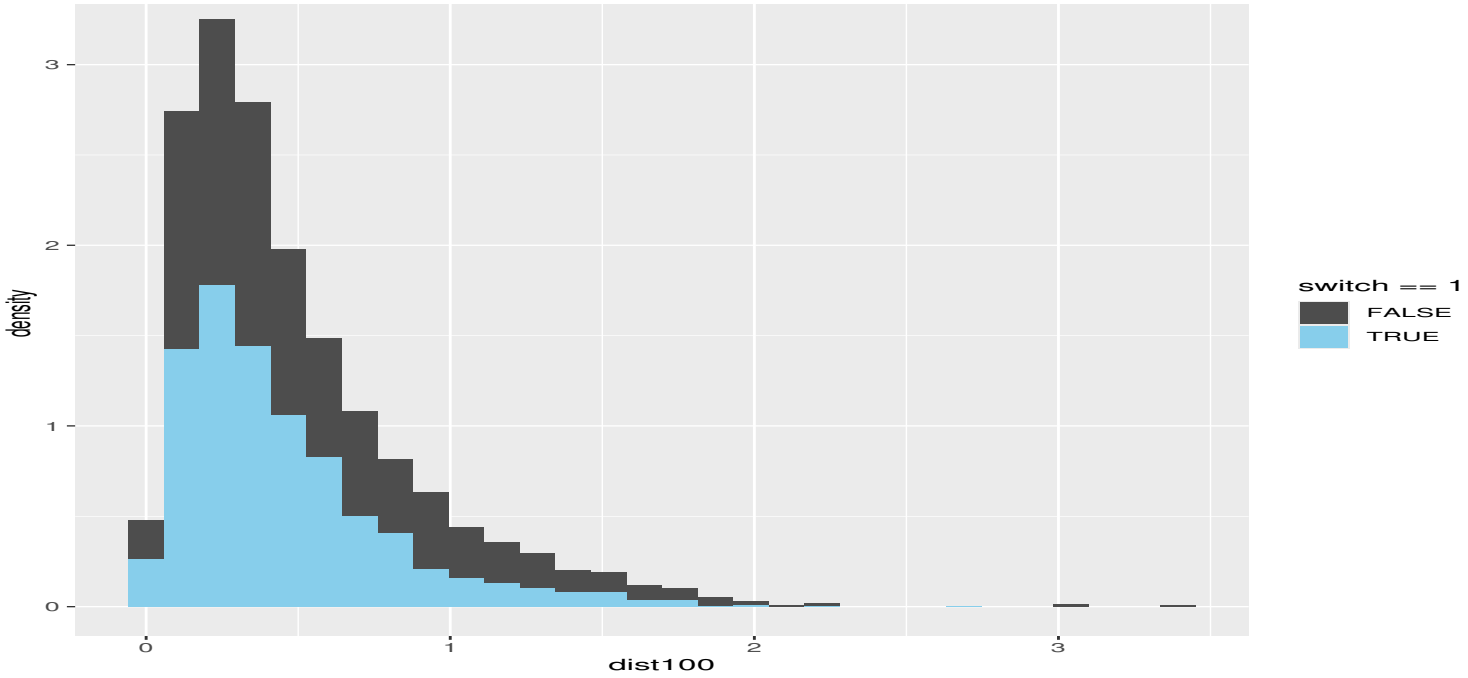
LOGISTIC REGRESSION

```
library(rstanarm)
data(wells)
wells$dist100 <- wells$dist / 100
head(wells)
library(ggplot2)
ggplot(wells,aes(x=dist100,y=after_stat(density),fill=switch==1)) +
geom_histogram() + scale_fill_manual(values=c("gray30", "skyblue"))
```

- Distribution of dist100: 1737 residents who switched (blue bars) and 1283 who did not (dark grey bars)

- We use a Student t prior with coefficients close to 0 but with chances of being large (less likely under Gaussian)

LOGISTIC REGRESSION



- It is just one density (not two!) which describes also the proportion of switch (blue) /no switch (dark grey) at various distances
- For the residents who switched wells, the distribution of `dist100` is more concentrated at smaller distances

LOGISTIC REGRESSION

```
t_prior <- student_t(df = 7, location = 0, scale = 2.5)
fit1<-stan_glm(switch ~ dist100,data=wells,seed = 12345,
  family = binomial(link = "logit"),
  prior = t_prior, prior_intercept = t_prior)
summary(fit1,digits=3)
round(posterior_interval(fit1, prob = 0.5), 3) # digits=3
fit2 <- update(fit1, formula = switch ~ dist100 + arsenic)
round(coef(fit2), 3)
summary(fit2,digits=3)
fit3<-stan_glm(switch ~ arsenic+assoc+educ+dist100,data=wells,
family = binomial(link = "logit"),seed = 12345,
  prior = t_prior, prior_intercept = t_prior)
  summary(fit3,digits=3)
```

LOGISTIC REGRESSION

- `switch` – binary/dummy (0 or 1) for well-switching
- **0.468**: `arsenic` – arsenic level in respondent's well
- **-0.897**: `dist100` – distance (100 meters) from the respondent's house to the nearest well with safe drinking water
- **-0.125**: `association` – binary/dummy (0 or 1) if member(s) of household participate in community organizations
- **0.043**: `educ` – years of education (head of household)
- Interpretation of those numbers (posterior means)?

LOGISTIC REGRESSION

Estimates:

	mean	sd	10%	50%	90%
(Intercept)	-0.157	0.103	-0.291	-0.157	-0.026
arsenic	0.468	0.042	0.414	0.468	0.522
assoc	-0.125	0.077	-0.223	-0.125	-0.026
educ	0.043	0.010	0.030	0.042	0.055
dist100	-0.897	0.107	-1.033	-0.896	-0.759

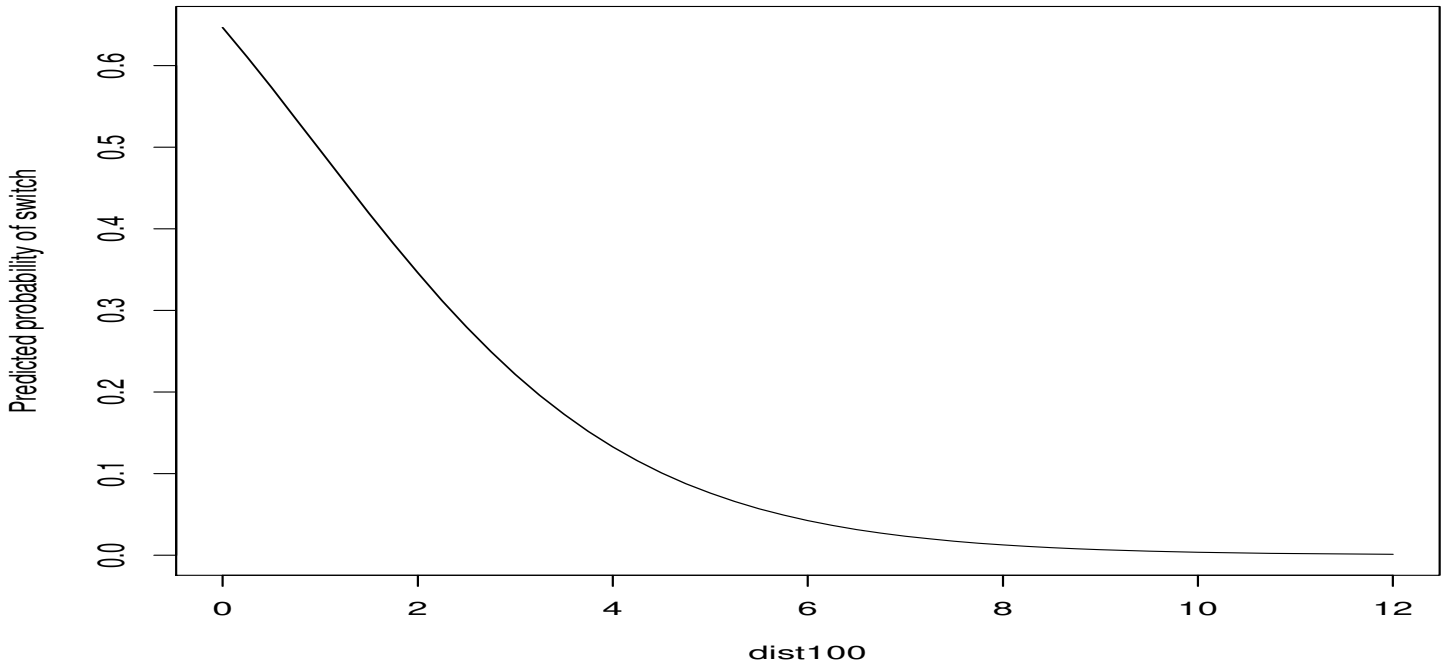
LOGISTIC REGRESSION

- Using the coefficient estimates from the first model, we can plot the predicted probability of `switch = 1` (as a function of `dist100`)

- `plogis` is the cdf of a logistic distribution

```
t_prior <- student_t(df = 7, location = 0, scale = 2.5)
fit1<-stan_glm(switch ~ dist100,data=wells,seed = 12345,
  family = binomial(link = "logit"),
  prior = t_prior, prior_intercept = t_prior)
summary(fit1,digits=3)
pr_switch <- function(x, ests) plogis(ests[1] + ests[2] * x)
coef(fit1)[1]; coef(fit1)[2]
aa=seq(0,12,0.25)
plot(aa,pr_switch(aa,coef(fit1)),type='l',xlab='dist100',
ylab='Predicted probability of switch')
```

LOGISTIC REGRESSION



Predicted probability of `switch = 1` as a function of `dist100`