Bayesian Statistics

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/

ALL BAYESIANS IN DAILY LIFE?

Interest in Milano or not?

- Prior knowledge
  - What is Milano? City, cookie, car?
  - Where is 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 $\pi(\lambda)$

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

EXERCISE: BAYES THEOREM

- Suppose a person is testing for diabetes
- A priori, the person has one chance out of a million of having diabetes
- In 3% of cases the test is positive although the person has no diabetes
  ($\Rightarrow$ *False positive error rate*)
- In 1% of cases the test is negative although the person has diabetes
  ($\Rightarrow$ *False negative error rate*)
- What is the probability that the person has diabetes when the test is positive?
- How does such probability change when a priori the patient has the same probability of having or not having diabetes?

BAYESIAN STATISTICS

Bayesian statistics is . . .

- . . . another way to make inference and forecast on population features (*practitioner's view*)
- . . . a way to learn from experience and improve own knowledge (*educated layman's view*)
- . . . a formal tool to combine prior knowledge and experiments (*mathematician's view*)
- . . . cheating (*hardcore frequentist statistician's view*)
- . . .

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

NOTIONS OF PROBABILITY

- Classical (random choice, equally likely events): Probability as $\dfrac{\#\text{Favourable events}}{\#\text{Possible events}}$
- Frequentist: Probability as asymptotic limit of frequency, i.e., of proportion of favourable events
- Subjective/Bayesian: Probability based on beliefs on, e.g., both head in tossing a coin (like previous) and final exam success (unlike previous)
- Axiomatic (Kolmogorov) on $(\Omega, \mathcal{F}, P)$, which contains the other three:
  - $P(A) \geq 0$ for all $A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for all mutually exclusive $A_i's \in \mathcal{F}$

Bayesian $\Rightarrow$ need to specify subjective P in $(\Omega, \mathcal{F}, P)$

ASSESSMENT OF PRIOR PROBABILITIES

T= person having a tumor in his/her life
I= person having an infarction in his/her life

$$\mathbb{P}(T \bigcup I) = .2,\ \mathbb{P}(T) = .3,\ \mathbb{P}(I) = .05,\ \mathbb{P}(T \bigcap I) = .1$$

ASSESSMENT OF PRIOR PROBABILITIES

T= person having a tumor in his/her life
I= person having an infarction in his/her life

$\mathbb{P}(T \bigcup I) = .2,\ \mathbb{P}(T) = .3,\ \mathbb{P}(I) = .05,\ \mathbb{P}(T \bigcap I) = .1$

- $\mathbb{P}(T \bigcup I) \geq \mathbb{P}(T)$

- $\mathbb{P}(I) \geq \mathbb{P}(T \bigcap I)$

ASSESSMENT OF PRIOR PROBABILITIES

T= person having a tumor in his/her life
I= person having an infarction in his/her life

$$\mathbb{P}(T \bigcup I) = .3,\ \mathbb{P}(T) = .2,\ \mathbb{P}(I) = .2,\ \mathbb{P}(T \bigcap I) = .15$$

ASSESSMENT OF PRIOR PROBABILITIES

T= person having a tumor in his/her life
I= person having an infarction in his/her life

$\mathbb{P}(T \bigcup I) = .3,\ \mathbb{P}(T) = .2,\ \mathbb{P}(I) = .2,\ \mathbb{P}(T \bigcap I) = .15$

- $.3 = \mathbb{P}(T \bigcup I) = \mathbb{P}(T) + \mathbb{P}(I) - \mathbb{P}(T \bigcap I) = .25$

- $\mathbb{P}(T \bigcup I) = .3,\ \mathbb{P}(T) = .2,\ \mathbb{P}(I) = .2,\ \mathbb{P}(T \bigcap I) = .1$

$\Rightarrow$ 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 $\Rightarrow P(A) = 1 - (364/365)^n$
- $$\begin{aligned} n = 10 &\quad\Rightarrow\quad P(A) = 0.027 \\ n = 50 &\quad\Rightarrow\quad P(A) = 0.128 \\ n = 100 &\quad\Rightarrow\quad P(A) = 0.240 \\ n = 200 &\quad\Rightarrow\quad P(A) = 0.422 \\ n = 300 &\quad\Rightarrow\quad P(A) = 0.561 \end{aligned}$$
- Therefore, what is your opinion about $P(A)$?

ASSESSING DISCRETE DISTRIBUTIONS: BETS

Probability Italy will win next FIFA World Cup

1. I bet $Y = 10\$$ on the Italian victory. How much are you willing to bet with me against the victory? (Say $10\$$ the first time, then $15\$$ and $20\$$)

2. Now let's reverse. You bet $Y = 10\$$ on the victory and you suggest my *fair* bet on the loss (Say $30\$$ the first time, then $25\$$ and $20\$$)

3. Let's repeat 1 and 2 until it is indifferent for you to bet either on the loss or the victory (i.e. $20\$$)

4. Let Y be the amount I bet on the victory of Italy

5. Let X be the amount you bet on the loss of Italy

6. Fair bet $\Rightarrow$ equal expected losses: $YP(loss) = XP(victory)$

7. $P(victory) = 1 - P(loss) \Rightarrow P(loss) = \dfrac{X}{X+Y} = \dfrac{20}{20+10} = \dfrac{2}{3}$

ASSESSING DISCRETE DISTRIBUTIONS: BETS

Problems

- Many people do not like to bet
- Most people dislike the idea of losing money
- I was talking about a $10\$$ bet, but would you have bet $1000X$ if I had bet $10,000\$$?
- Reaching convergence to a fair bet might be a long process

REFERENCE LOTTERIES

1. Lottery 1
   - Get a trip to Australia if Italy wins
   - Stay at home if Italy looses
2. Lottery 2
   - Get a trip to Australia with probability p, e.g. if a random number generated from a uniform distribution on $[0, 1]$ is $\leq p$
   - Stay at home with probability $1 - p$, e.g. if a random number generated from a uniform distribution on $[0, 1]$ is $> p$
3. Specify p_1. Which lottery do you prefer?
4. If Lottery 1 is preferred offer change p_i to $p_{i+1} > p_i$.
5. If Lottery 2 is preferred offer change p_i to $p_{i+1} < p_i$.
6. When indifference point is reached $\Rightarrow P(victory) = p_i$, else Goto 4.

ASSESSING CONTINUOUS DISTRIBUTIONS

X continuous random variable (e.g. light bulb lifetime)

- Choose $x_1, \ldots, x_n$
- Assess $F(x_i) = P(X \leq x_i), i = 1, n$
- Draw $F(x)$
- Look at $F(x)$ at some points for consistency

or

- Choose probabilities $p_1, \ldots, p_n$
- Find x_i's s.t. $F(x_i) = P(X \leq x_i) = p_i, i = 1, n$
- Draw $F(x)$
- Look at $F(x)$ at some points for consistency

BAYES THEOREM AND LIKELIHOOD

- Sample $\underline{X} = (X_1, \ldots, X_n)$, i.i.d. from $f(x|\lambda) \Rightarrow$ likelihood $l_x(\lambda) = \prod_{i=1}^n f(X_i|\lambda)$

- Prior $\pi(\lambda) \Rightarrow$ posterior $\pi(\lambda|\underline{X}) = \dfrac{l_x(\lambda)\pi(\lambda)}{\int l_x(\theta)\pi(\theta)d\theta}$

- I.i.d. property not necessarily needed to get likelihood, e.g. Markovian observations where $f(X_1, \ldots, X_n|\lambda) = f(X_1|\lambda)\prod_{i=2}^n f(X_i|X_{i-1}, \lambda)$

- The likelihood is all that we need from data to perform inference and, given it, the way the experiment was performed is not relevant (*Likelihood Principle*)
  - Compare two experiments counting the number x of heads in n tosses of a coin knowing that $P(head) = \theta$
  - The sequence $HHT\ldots TH$ is known $\Rightarrow \theta^x(1-\theta)^{n-x}$
  - Only known about x heads and $n-x$ tails $\Rightarrow \binom{n}{x}\theta^x(1-\theta)^{n-x}$
  - Different probabilities but $\theta^x(1-\theta)^{n-x}$ is the same contribution to the likelihood

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

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

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: $\dfrac{\sum_{i=1}^{n}X_i/\sigma^2+\mu_0/\tau^2}{n/\sigma^2+1/\tau^2}$

CONJUGATE PRIOR FOR GAUSSIAN

- Lack of knowledge about μ given by noninformative prior
- $\pi(\mu) \propto c$, c positive constant
- What is the problem with this prior?

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

- $\Rightarrow \mu|\underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n}X_i}{n}, \frac{\sigma^2}{n}\right)$
- Posterior mean = MLE = $\frac{\sum_{i=1}^{n}X_i}{n}$

JEFFREYS PRIORS

- There are alternative proper noninformative priors:
  - Flat prior on $[-K, K], K > 0$: $\pi(\mu) = \frac{1}{2K} I_{[-K,K]}(\mu)$ (I_A indicator function of set A)
  - Diffuse prior: $\mu \sim \mathcal{N}(\mu_0, 10^6)$
- The previous prior $\pi(\mu) \propto c$ is an example of Jeffreys priors
- $\underline{\theta} = (\theta_1, \ldots, \theta_p)$ p-dimensional parameter in $f(X|\underline{\theta})$
- $J = \{J_{ij}\}_{i,j=1,\ldots,p}$ Fisher information matrix s.t. for each i, j

$$\begin{aligned} J_{ij} &= -E\left[\frac{\partial^2 \log(f(X|\underline{\theta}))}{\partial\theta_i \partial\theta_j}\right] \\ &= E\left[\left(\frac{\partial \log(f(X|\underline{\theta}))}{\partial\theta_i}\right)\left(\frac{\partial \log(f(X|\underline{\theta}))}{\partial\theta_j}\right)\right] \end{aligned}$$

JEFFREYS PRIORS FOR GAUSSIAN

- Jeffreys prior: $\pi(\theta) \propto \sqrt{|J|}$, with $|J|$ the determinant of the Fisher information matrix
- Gaussian model with known variance and unknown mean μ
- Here the matrix is of size 1 since there is just one parameter

$$
\begin{aligned}
\pi(\mu) &\propto \sqrt{|J|} \propto \sqrt{E\left(\frac{\partial \log(f(X|\mu))}{\partial \mu}\right)^2} \\
&\propto \sqrt{E\left(\frac{X-\mu}{\sigma^2}\right)^2} \propto \frac{1}{\sigma^2}\sqrt{\int f(X|\mu)\,(X-\mu)^2\,dX} \\
&\propto \frac{\sigma}{\sigma^2} \propto \frac{1}{\sigma} \propto 1
\end{aligned}
$$

- The last step is possible since σ^2 is a constant here

CONJUGATE PRIOR FOR GAUSSIAN

- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
- Mean/median $\mu \in \Re$ and variance $\sigma^2 > 0$ unknown
- $\underline{X} = (X_1, \ldots, X_n)$
- Conjugate normal-inverse gamma prior
- Prior $\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2)$
- $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \tau^2\sigma^2)$
- $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ Inverse gamma
- $\pi(\sigma^2) = \dfrac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}$, with $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$

CONJUGATE PRIOR FOR GAUSSIAN

- After some computations (left as an exercise) we get the posterior $\pi(\mu, \sigma^2 | \underline{X}) = \pi(\mu | \sigma^2, \underline{X}) \pi(\sigma^2 | \underline{X})$ with
  - $\mu | \sigma^2, \underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n} X_i + \mu_0/\tau^2}{n + 1/\tau^2}, \frac{\sigma^2}{n + 1/\tau^2}\right)$
  - $\sigma^2 | \underline{X} \sim \mathcal{IG}(\alpha + (n+1)/2, \beta)$

- The posterior marginal of μ, i.e. $\pi(\mu | \underline{X})$, has a Student-t distribution

- $\sigma^2 | \mu, \underline{X} \sim \mathcal{IG}(\alpha + (n+1)/2, \beta + \sum_{i=1}^{n} (X_i - \mu)^2/2 + (\mu - \mu_0)^2/(2\tau^2))$

- $\Rightarrow$ useful for MCMC (Gibbs sampling)

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 $L(\lambda, a) = (\lambda - a)^2$

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

QUICK GLIMPSE TO BAYESIAN DECISION ANALYSIS

- Bayesian Decision Analysis supports a Decision Maker in making decisions under uncertainty:
  - Set of alternatives (actions) $a \in \mathcal{A}$
  - Unknown parameter θ depending on *state of nature*
  - Consequence $c(a, \theta)$ of action a when θ occurs
  - Loss function $L(c(a, \theta))$
  - Posterior distribution $\pi(\theta|x)$ on parameter θ, after observing x
  - Optimal action satisfies the Minimum (Subjective) Expected Loss Principle:

$$a^* = \arg\min_{a \in \mathcal{A}} \int L(c(a, \theta))\pi(\theta|x)d\theta$$

- Often losses are replaced by utilities and minimisation becomes maximisation

QUICK GLIMPSE TO BAYESIAN DECISION ANALYSIS

- State of nature: $\theta = \{\text{Rain today, No rain today}\}$
- Actions $a = \{\text{stay at home, go out with umbrella, go out without umbrella}\}$
- Consequences $c(a, \theta)$, e.g., $c(\text{stay at home, No rain today}) =$ fired at work or $c(\text{go out without umbrella, Rain today}) =$ unable to meet an important customer
- Loss function $L(c(a, \theta))$, e.g., $L(c(\text{stay at home, No rain today})) = 100,000$ (income loss, in euros, after being fired)
- Posterior distribution $\pi(\theta|x)$ on parameter θ, after observing x, e.g., rain in the previous days or weather forecasts
- Optimal action (suppose *go out with umbrella*) satisfies the Minimum (Subjective) Expected Loss Principle:

$$a^* = \arg\min_{a \in \mathcal{A}} \int L(c(a, \theta))\pi(\theta|x)d\theta$$

PARAMETER ESTIMATION

- Light bulb: posterior mean $\hat{\lambda} = \dfrac{\alpha + n}{\beta + \sum_{i=1}^{n} X_i}$

  $\Rightarrow$ compare with

  - prior mean $\dfrac{\alpha}{\beta}$
  - MLE $\dfrac{n}{\sum_{i=1}^{n} X_i}$

- MAP (Maximum a posteriori)

  $\Rightarrow \hat{\lambda} = \dfrac{\alpha + n - 1}{\beta + \sum X_i}$

PRIOR AND DATA INFLUENCE

- 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)

EXERCISE: PARAMETER ESTIMATION

Prior influence (multinomial data and Dirichlet prior)

$(n_1, \ldots, n_k) \sim \mathcal{MN}(n, p_1, \ldots, p_k)$

$(p_1, \ldots, p_k) \sim \mathcal{D}ir(s\alpha_1, \ldots, s\alpha_k),\ \sum \alpha_i = 1,\ s > 0$

- Posterior mean: $p_i^* = \dfrac{s\alpha_i + n_i}{s + n}$
- Prior mean: $\tilde{p}_i = \alpha_i$
- MLE: $\dfrac{n_i}{n}$
- $s \to 0 \Rightarrow p_i^* \to$ MLE
- $s \to \infty \Rightarrow p_i^* \to \tilde{p}_i$

PRIOR CHOICE

Where to start from?

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

$\Rightarrow$ *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. $P(X_i \leq x_q) = q$
- Quantiles of λ, i.e. $P(\lambda \leq \lambda_q) = q$
- Moments $\mathbf{E}\lambda^k$ of λ, i.e. $\int \lambda^k \pi(\lambda) d\lambda = a_k \Leftrightarrow \int (\lambda^k - a_k)\pi(\lambda) d\lambda = 0$
- Generalised moments of λ, i.e. $\int h(\lambda)\pi(\lambda) d\lambda = 0$
- Most likely value and upper and lower bounds
- $\ldots$
- None of them

PRIOR CHOICE

How to get information?

- 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)

- Split of possible values of λ or X_i into equally likely intervals $\Rightarrow$ quantiles

- Most likely value and upper and lower bounds

- *Expected* value of λ and *confidence* on such value (mean and variance)

- ...

PRIOR CHOICE

Which prior?

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

PRIOR CHOICE

Choice of a prior

- Defined on suitable set (interval vs. positive real)
- Suitable functional form (monotone/unimodal, heavy/light tails, etc.)
- Mathematical convenience
- *Tradition* (e.g. lognormal for engineers)

PRIOR CHOICE

Gamma prior - choice of hyperparameters

- $X_1, \ldots, X_n \sim \mathcal{E}(\lambda)$
- $f(X_1, \ldots, X_n | \lambda) = \lambda^n \exp\{-\lambda \sum X_i\}$
- $\lambda \sim \mathcal{G}(\alpha, \beta) \Rightarrow f(\lambda | \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} \exp\{-\beta\lambda\} / \Gamma(\alpha)$
- $\Rightarrow \lambda | X_1, \ldots, X_n \sim \mathcal{G}(\alpha + n, \beta + \sum X_i)$

PRIOR CHOICE

Gamma prior - choice of hyperparameters

- $\mathcal{E}\lambda = \mu = \alpha/\beta$ and $Var\lambda = \sigma^2 = \alpha/\beta^2$
  $\Rightarrow \alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$

- Two quantiles $\Rightarrow (\alpha, \beta)$ using, say, Wilson-Hilferty approximation. Third quantile specified to check consistency

- *Hypothetical experiment*: posterior $\mathcal{G}(\alpha + n, \beta + \sum X_i)$
  $\Rightarrow \alpha$ *sample size* and β *sample sum*

BAYESIAN SIMULATIONS

Alternative choice: $\lambda \sim \mathcal{LN}(\alpha, \beta)$

- no posterior in closed form $\Rightarrow$ numerical simulation

Markov Chain Monte Carlo (MCMC):

- draw$^{(*)}$ a sample $\lambda^{(1)}, \lambda^{(2)}, \ldots$ (Monte Carlo) $\ldots$
- $\ldots$ from a Markov Chain whose stationary distribution is $\ldots$
- $\ldots$ the posterior $\pi(\lambda|\underline{X})$ and compute $\ldots$
- $\mathcal{E}(\lambda|\underline{X}) \approx \sum_{i=m+1}^{n} \lambda^{(i)}/(n-m)$, etc.

(*) For $\lambda = (\theta, \mu) \Rightarrow$ Gibbs sampler:

- draw $\theta^{(i)}$ from $\theta|\mu^{(i-1)}, \underline{X}$
- draw $\mu^{(i)}$ from $\mu|\theta^{(i)}, \underline{X}$
- repeat *until convergence*

MCMC: REGRESSION

- $y = \beta_0 + \beta_1 x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $(y_1, x_1), \ldots, (y_n, x_n)$
- Likelihood $\propto (\sigma^2)^{-n/2} \exp\{\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\}$
- Priors: $\beta_0 \sim \mathcal{N}, \beta_1 \sim \mathcal{N}, \sigma^2 \sim \mathcal{IG}$
- Full posterior conditionals:
  - $\beta_0 | \beta_1, \sigma^2 \sim \mathcal{N}$
  - $\beta_1 | \beta_0, \sigma^2 \sim \mathcal{N}$
  - $\sigma^2 | \beta_0, \beta_1 \sim \mathcal{IG}$

  $\Rightarrow$ MCMC

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%

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

- Two sided test: $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$
  - 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 $\mathbb{P}(\lambda_0|\underline{X})$

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

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 $\Gamma(n+1) = n\Gamma(n)$

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

$$\Rightarrow \frac{P(\mathcal{M}_1|data)}{P(\mathcal{M}_2|data)} = \frac{P(data|\mathcal{M}_1)}{P(data|\mathcal{M}_2)} \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \cdot \frac{1/P(data)}{1/P(data)} = BF \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}$$

BACK TO HYPOTHESIS TESTING

- $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, with $\Theta = \Theta_0 \bigcup \Theta_1$
- $\pi_0(\theta)$ prior on Θ_0 and $\pi_1(\theta)$ prior on Θ_1
- Priors on hypotheses: $P(\Theta_0) = \varepsilon$ and $P(\Theta_1) = 1 - \varepsilon$
- Mixture prior on Θ: $\pi_\varepsilon(\theta) = \varepsilon\pi_0(\theta)I_{\Theta_0}(\theta) + (1-\varepsilon)\pi_1(\theta)I_{\Theta_1}(\theta)$
- $I_A(x)$ indicator function of set A
- Likelihood $l_x(\theta) = f(\underline{X}|\theta)$
- Posterior $\pi_\varepsilon(\theta|\underline{X}) = \dfrac{\varepsilon l_x(\theta)\pi_0(\theta)I_{\Theta_0}(\theta) + (1-\varepsilon)l_x(\theta)\pi_1(\theta)I_{\Theta_1}(\theta)}{\varepsilon\int_{\Theta_0} l_x(\theta)\pi_0(\theta)d\theta + (1-\varepsilon)\int_{\Theta_1} l_x(\theta)\pi_1(\theta)d\theta}$

BACK TO HYPOTHESIS TESTING

- Posterior on hypotheses
  - $P(\Theta_0|\underline{X}) = \dfrac{\varepsilon \int_{\Theta_0} l_x(\theta)\pi_0(\theta)d\theta}{\varepsilon \int_{\Theta_0} l_x(\theta)\pi_0(\theta)d\theta + (1-\varepsilon)\int_{\Theta_1} l_x(\theta)\pi_1(\theta)d\theta}$
  - $P(\Theta_1|\underline{X}) = \dfrac{(1-\varepsilon)\int_{\Theta_1} l_x(\theta)\pi_1(\theta)d\theta}{\varepsilon \int_{\Theta_0} l_x(\theta)\pi_0(\theta)d\theta + (1-\varepsilon)\int_{\Theta_1} l_x(\theta)\pi_1(\theta)d\theta}$
- Posterior odds = Bayes factor · prior odds
- $\dfrac{P(\Theta_0|\underline{X})}{P(\Theta_1|\underline{X})} = \dfrac{\int_{\Theta_0} l_x(\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} l_x(\theta)\pi_1(\theta)d\theta} \cdot \dfrac{\varepsilon}{1-\varepsilon}$
- Posterior odds influenced by prior odds, i.e. choice of prior on hypotheses
- $\Rightarrow$ Often only Bayes factor is used in hypothesis testing (corresponds to $\varepsilon = 0.5$)

PRIORS AND MODELS

- The Bayesian approach criticized because *subjective* but . . .

- . . . is the choice of the model (the only aspect which matters in the frequentist approach) really *objective*?

- Consider the failure times of n cars:

- $\left\{X_{ij_i}\right\}, i = 1, \ldots, n; j_i = 1, \ldots, n_i$

- Who is choosing the model? Expert and statistician, like for the prior!

MODEL SELECTION

Before the analysis - Model chosen according to

- physical laws
- mathematical convenience
- exploratory data analysis
  - Weibull plot, Duane plot, q-q plot
  - histogram
- our knowledge about experiment, e.g.
  - same/similar/different car and same/different cause of failure?
  - replacement policy and aging
- ...

MODEL SELECTION

Which model for $\{X_{ij_i}\}, i = 1, \ldots, n; j_i = 1, \ldots, n_i$?

- All the cars behave in the same way and the failure pattern is not changing over time $\Rightarrow X_{ij_i} \sim \mathcal{E}(\lambda)$
- The cars behave differently and the failure pattern is not changing over time $\Rightarrow X_{ij_i} \sim \mathcal{E}(\lambda_i)$
- All the cars behave in the same way and the failure pattern is changing over time $\Rightarrow X_{ij_i}$ from a NHPP (Nonhomogeneous Poisson process) with intensity $\lambda(t)$
- The cars behave differently and the failure pattern is not over time $\Rightarrow X_{ij_i}$ from NHPP's with intensities $\lambda_i(t)$
- Each failure affects only the next one (Markov property, e.g. $AR(1)$ model) $\Rightarrow X_{i,k+1} = \rho X_{i,k} + \varepsilon_{i,k}$
- etc.
- Lognormal, Weibull, Birnbaum-Saunders, etc. instead of exponential

MODEL SELECTION

After the analysis - Model chosen according to

- graphical displays (e.g. residuals in regression)
- goodness of fit tests (e.g. χ^2, Kolmogorov-Smirnov) (*not very Bayesian!*)
- Bayes factor to compare
  $\mathcal{M}_1 = \{f_1(x|\theta_1), \pi(\theta_1)\}$ and $\mathcal{M}_2 = \{f_2(x|\theta_2), \pi(\theta_2)\}$

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

- Posterior odds

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

- AIC, BIC, DIC et al.

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
- Alternative: Posterior median w.r.t. posterior mean

BAYESIAN ROBUSTNESS

- Practical impossibility of specifying priors exactly matching experts' knowledge
- Prior elicitation subject to uncertainty and, possibly, some degree of arbitrariness introduced by the analyst, e.g. the functional form of the distribution
- Uncertainty in the choice of priors modelled through a class of distribution (the same might apply for loss functions and statistical models/likelihoods)
- Use of indices to measure the consequences (i.e. perform robustness analysis) of the choice of a class of priors on the quantities of interest (e.g. posterior mean)
- An answer to the criticism about the arbitrariness in the choice of the prior and a possible excessive influence

BAYESIAN ROBUSTNESS

A more formal statement about model and prior sensitivity

- $M = \{Q_\theta; \theta \in \Theta\}$, Q_θ probability on $(\mathcal{X}, \mathcal{F}_\mathcal{X})$
- Sample $\underline{x} = (x_1, \ldots, x_n) \Rightarrow$ likelihood $l_x(\theta) \equiv l_x(\theta | x_1, \ldots, x_n)$
- Prior P su $(\Theta, \mathcal{F}) \Rightarrow$ posterior P^*
- Uncertainty about M and/or $P \Rightarrow$ changes in
  - $E_{P^*}[h(\theta)] = \dfrac{\int_\Theta h(\theta) l(\theta) P(d\theta)}{\int_\Theta l(\theta) P(d\theta)}$
  - P^*

Bayesian robustness studies these changes

ROBUST BAYESIAN ANALYSIS

Interest in robustness w.r.t. to changes in prior/model/loss but most work concentrated on priors since

- controversial aspect of Bayesian approach
- easier (w.r.t. model) computations
- problems with interpretation of classes of models/likelihood
- often interest in posterior mean (corresponding to optimal Bayesian action under squared loss function) and no need for classes of losses

ROBUST BAYESIAN ANALYSIS

Three major approaches

- *Informal sensitivity*: comparison among few priors
- *Global sensitivity*: study over a class of priors specified by some features
- *Local sensitivity*: infinitesimal changes w.r.t. elicited prior

ROBUST BAYESIAN ANALYSIS

We concentrate mostly on sensitivity to changes in the prior

- Choice of a class Γ of priors
- Computation of a robustness measure, e.g. range $\delta = \overline{\rho} - \underline{\rho}$ $(\overline{\rho} = \sup_{P \in \Gamma} E_{P*}[h(\theta)]$ and $\underline{\rho} = \inf_{P \in \Gamma} E_{P*}[h(\theta)])$
  - δ "small" $\Rightarrow$ robustness
  - δ "large", $\Gamma_1 \subset \Gamma$ and/or new data
  - δ "large", Γ and same data

ROBUST BAYESIAN ANALYSIS

Relaxing the unique prior assumption (Berger and O'Hagan, 1988)

- $X \sim \mathcal{N}(\theta, 1)$
- Prior $\theta \sim \mathcal{N}(0, 2)$
- Data $x = 1.5 \Rightarrow$ posterior $\theta | x \sim \mathcal{N}(1, 2/3)$
- Split $\Re$ in intervals with same probability p_i as prior $\mathcal{N}(0, 2)$

ROBUST BAYESIAN ANALYSIS

Refining the class of priors (Berger and O'Hagan, 1988)

I_i	p_i	p_i^*	Γ_Q	Γ_{QU}
(-∞,-2)	0.08	.0001	(0,0.001)	(0,0.0002)
(-2,-1)	0.16	.007	(0.001,0.029)	(0.006,0.011)
(-1,0)	0.26	.103	(0.024,0.272)	(0.095,0.166)
(0,1)	0.26	.390	(0.208,0.600)	(0.322,0.447)
(1,2)	0.16	.390	(0.265,0.625)	(0.353,0.473)
(2,+∞,)	0.08	.110	(0,0.229)	(0,0.156)

- Γ_Q quantile class and Γ_{QU} unimodal quantile class
- Robustness in Γ_{QU}
- Huge reduction of δ from Γ_Q to Γ_{QU}

CLASSES OF PRIORS

Desirable features of classes of priors

- Easy elicitation and interpretation (*e.g. moments, quantiles, symmetry, unimodality)*
- Compatible with prior knowledge (*e.g. quantile class)*
- Simple computations
- Without unreasonable priors (*e.g. unimodal quantile class, ruling out discrete distributions)*

CLASSES OF PRIORS

- $\Gamma_P = \{P : p(\theta; \omega), \omega \in \Omega\}$ *(Parametric class)*
  - $\Gamma_P = \{\mathcal{G}(\alpha, \beta) : l_1 \le \alpha/\beta \le u_1, l_2 \le \alpha/\beta^2 \le u_2\}$
- $\Gamma_Q = \{P : \alpha_i \le P(I_i) \le \beta_i, i = 1, \ldots, m\}$ *(Quantile class)*
- $\Gamma_{QU} = \{P \in \Gamma_Q,$ unimodal$\}$ *(Unimodal quantile class)*
- $\Gamma_{QUS} = \{P \in \Gamma_{QU},$ symmetric$\}$ *(Symmetric, unimodal quantile class)*

CLASSES OF PRIORS

- $\Gamma_{GM} = \{P : \int h_i(\theta)dP(\theta) = a_i, i = 1, \ldots, m\}$ *(Generalised moments class)*
  - $h_i(\theta) = \theta^i$ *(Moments class)*
  - $h_i(\theta) = I_{A_i}(\theta)$ *(Quantile class)*
- $\Gamma^B = \{P : L(\theta) \leq p(\theta) \leq U(\theta)\}$ *(Density bounded class)*
- $\Gamma^{DB} = \{F \text{ c}.d.f. : F_l(\theta) \leq F(\theta) \leq F_u(\theta), \forall\theta\}$ *(Distribution bounded class)*