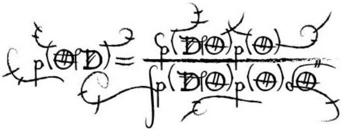


Bayesian Learning 732A46: Lecture 1

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

March 2015

Course overview

- A few words about me
  - M.Sc. Engineering Mathematics, Lund University
  - Ph.D. Statistics, 2015, Stockholm University
    Supervisor: **Mattias Villani**
  - Doctoral thesis on **Markov Chain Monte Carlo** for large data sets.
  - I am a Bayesian believer.
- The course consist of **12 lectures** and **4 computer labs**.
  Material: https://github.com/matiasq/BayesLearningLiU
- Divided into four **modules** (**3** lectures + **1** lab each)
  1. The **basics**, single and multiparameter models.
  2. **Regression** models.
  3. Estimating complex models with **MCMC**.
  4. **Flexible models** and **Model inference**.
- **Examination**
  - Lab reports (2 credits, work in pairs).
  - An individual project with a written report (4 credits).
  - Oral exam (if needed).

Lecture overview

- The Bayesian paradigm
- The likelihood function
- The Bernoulli model
- The normal model with known variance

What is a statistical model?

- **Briefly**: A model is a **compact** and **interpretable** representation of the observed data.
- **Elements** of a statistical model
  - **Data** $y = (y_1, \ldots, y_n)$.
  - **Parameter(s)** θ.
  - A **probabilistic** model $p(y|\theta)$ - probability theory to represent **the uncertainty** that is inherent in data (noise, natural variation).
- **Learn** about (*the unknown*) θ.
- A statistician deals with **the uncertainty** regarding θ. True **regardless if she is a Frequentist or Bayesian**!
- The difference is how she **thinks about** this uncertainty...

Two different minds...

Reading the mind of a **Frequentist** statistician

I think of θ as an *unknown* but **non-random** **"state of nature"** (fixed quantity). The **random data** y are generated under this **fixed** θ via the model $p(y|\theta)$. I **could have** obtained **another dataset**, so I will use **a repeated sampling argument** to describe my uncertainty about θ.

Reading the mind of a **Bayesian** statistician

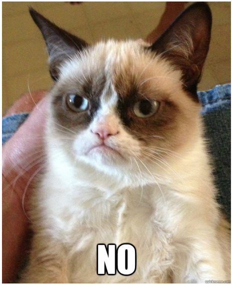
There are two quantities present; the data y and the **"state of nature"** θ. **I have seen** y but I have **not seen the unknown** θ. I will therefore regard θ as random and describe my uncertainty about θ **conditional on the data I have seen**.

The likelihood function

- The notation $p(y|\theta)$ for representing the **probabilistic model** is interpreted in two **distinct** ways.
- As a **FUNCTION OF** y, for a **FIXED** θ, $p(y|\theta)$ is the **probability distribution** for the data $y = (y_1, \ldots, y_n)$. $\int p(y|\theta)dy = 1$
- As a **FUNCTION OF** θ, for a **FIXED** y, $p(y|\theta)$ is the **likelihood function** for the parameter θ.
- **Question**: Is the **likelihood function** a probability distribution for θ?

The likelihood function

- The notation $p(y|\theta)$ for representing the **probabilistic model** is interpreted in two **distinct** ways.
- As a **FUNCTION OF** y, for a **FIXED** θ, $p(y|\theta)$ is the **probability distribution** for the data $y = (y_1, \ldots, y_n)$. $\int p(y|\theta)dy = 1$
- As a **FUNCTION OF** θ, for a **FIXED** y, $p(y|\theta)$ is the **likelihood function** for the parameter θ.
- **Question**: Is the **likelihood function** a probability distribution for θ?



Inversion of probabilities: Bayes' theorem

- Let H and E be two events.
- From a basic course in probability: **Bayes' theorem** relates $\Pr(H|E)$ to $\Pr(E|H)$

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}, \quad \Pr(E) = \Pr(E|H)\Pr(H) + \Pr(E|H^c)\Pr(H^c).$$

- For the **inference problem**

$$\begin{aligned} E &= \text{Evidence: data } y \\ H &= \text{Hypothesis about } \theta \text{ (e.g.: parameter, prediction).} \end{aligned}$$

- **Bayes' theorem** for the inference problem (continuous θ)

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \quad p(y) = \int p(y|\theta)p(\theta)d\theta.$$

- $p(\theta)$ is the prior distribution.
- $p(\theta|y)$ is a function of θ with y **regarded as fixed**.
- **Question**: Is $p(\theta|y)$ a probability distribution for θ?

The cat is not grumpy anymore!

Bayesian inference

- $p(\theta|y)$ is the **posterior distribution**. A statement like

$$\Pr(\theta \in [a, b]|y) \quad \text{makes sense.}$$ **Fantastic!**

In a world of classical statistics: A difficult question

Let $[a, b]$ be a (classical) confidence interval with significance $\alpha = 0.05$. **Conditional on** y (i.e. given that we have seen the data), what is the probability that $[a, b]$ covers θ?

Bayesian inference

- $p(\theta|y)$ is the **posterior distribution**. A statement like

$$\Pr(\theta \in [a, b]|y) \quad \text{makes sense.}$$ **Fantastic!**

In a world of classical statistics: A difficult question

Let $[a, b]$ be a (classical) confidence interval with significance $\alpha = 0.05$. **Conditional on** y (i.e. given that we have seen the data), what is the probability that $[a, b]$ covers θ?

Answer: 0 or 1! **Why?:** The interval $[a, b]$ is regarded as stochastic w.r.t the data y. Once the data is observed, there is no uncertainty anymore. Therefore, a **frequentist does not condition on the observed data**, but instead **averages over all possible data** that could have been observed (but were not observed!).

- **Punchline 1:** Bayesian inference is **conditional** on observed data, whereas classical inference is **unconditional** (averages over unobserved data).
- Bayesian inference obeys the **Likelihood principle**.

Bayesian inference, cont.

- **Punchline 2:** Confidence intervals are **hard to interpret** because they are **not probabilities** w.r.t θ. The **Bayesian posterior** is straightforward.
- Revisiting the formula

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \quad p(y) = \int p(y|\theta)p(\theta)d\theta.$$

  - **The prior** $p(\theta)$: your **subjective belief** about the uncertainty of θ.
  - **The likelihood** $p(y|\theta)$: the **information** about θ contained in y.
  - **The marginal likelihood** $p(y)$: A normalizing constant **independent** of θ.
- **Compact form**

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior.} \end{aligned}$$

Bayesian inference, cont.

- What if θ is a natural constant? **Example:** the speed of light.
- **Bayesian:** Do you know the value of θ or not?
- To a **Bayesian**, any unknown quantity is a random variable.
- **Subjective probability:** $p(\theta)$ reflects Your knowledge/**uncertainty** about θ.
- **Bayes' theorem:** Updates Your **subjective prior belief** **objectively** (just maths!) to a **posterior belief** by combining it with the data via the **likelihood function**.
- A probability distribution for θ is **useful for decision making**.

Bayes in action: Bernoulli model with a Beta prior

- **Model**

$$y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Bern}(\theta)$$

- **Prior** with *hyper parameters* α_0 and β_0

$$p(\theta) = \mathrm{Beta}(\theta|\alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} \text{ for } 0 \leq \theta \leq 1.$$

  α_0 and β_0 are **set by the user to reflect her uncertainty** about θ.

- **Posterior** [$s = \sum_{i=1}^n y_i$ nbr of successes, $f = n - s$]

$$\begin{aligned} p(\theta|y_1, ..., y_n) &\propto p(y_1, ..., y_n|\theta)p(\theta) \\ &\propto \theta^s(1-\theta)^f\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} \\ &= \theta^{s+\alpha_0 - 1}(1-\theta)^{f+\beta_0 - 1}. \end{aligned}$$

- This is **proportional to** the $\mathrm{Beta}(\theta|\alpha + s, \beta + f)$ density.
- The **prior-to-posterior** mapping reads

$$\theta \sim \mathrm{Beta}(\alpha_0, \beta_0) \overset{y_1, \ldots, y_n}{\Longrightarrow} \theta|y_1, ..., y_n \sim \mathrm{Beta}(\underbrace{\alpha_0 + s}_{\alpha_n}, \underbrace{\beta_0 + f}_{\alpha_n}).$$

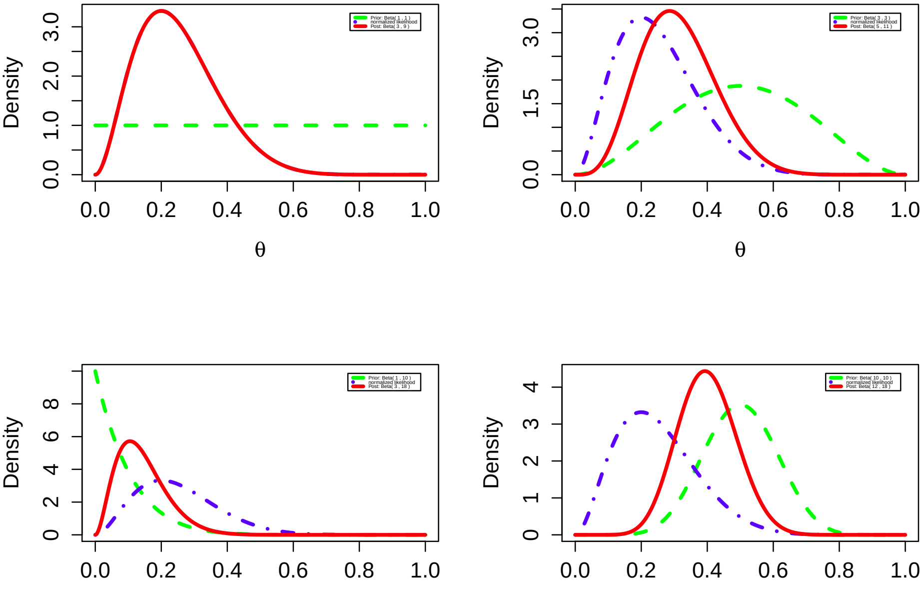
Bernoulli model: spam emails

- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam (and 2788 non-spam).

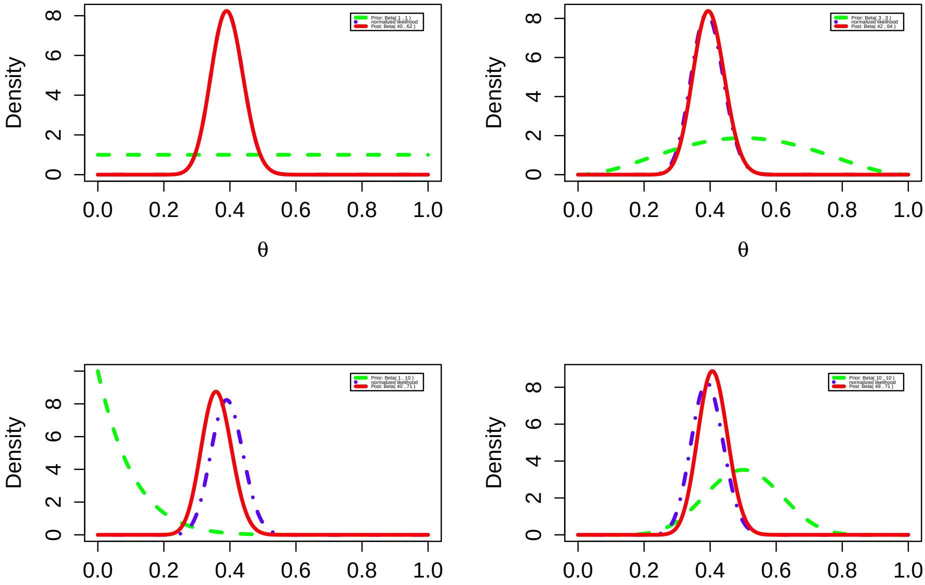
- Let $y_i = 1$ if ith email is spam (0 otherwise). Assume $y_i|\theta \overset{iid}{\sim} \mathrm{Bern}(\theta)$ and $\theta \sim \mathrm{Beta}(\alpha_0, \beta_0)$ **a priori**.

- **Posterior**

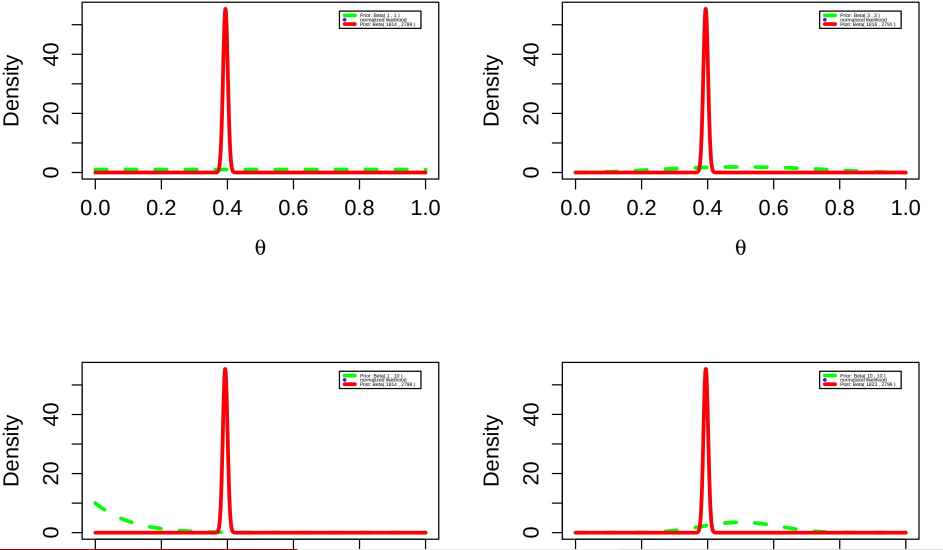
$$\theta|y \sim \mathrm{Beta}(\alpha_0 + 1813, \beta_0 + 2788).$$


Prior: Beta(1 , 1)
normalized likelihood
Post: Beta(3 , 9)
Density
θ
Prior: Beta(3 , 3)
normalized likelihood
Post: Beta(5 , 11)
Density
θ
Prior: Beta(1 , 10)
normalized likelihood
Post: Beta(3 , 18)
Density
Prior: Beta(10 , 10)
normalized likelihood
Post: Beta(12 , 18)
Density

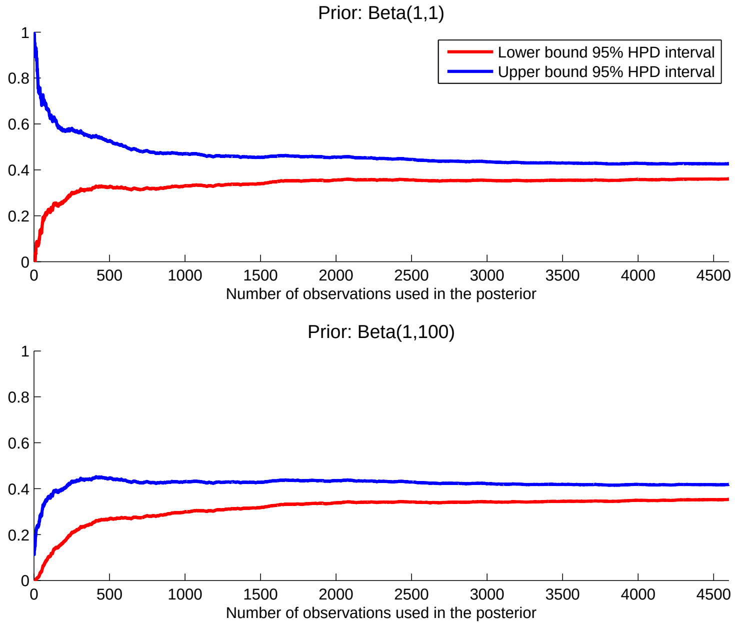
Spam data ($n = 100$, $s = 39$ and $s = 61$): prior sensitivity



Spam data ($n = 4601$, $s = 1813$ and $s = 2788$): prior sensitivity



Spam data: Posterior convergence



Normal model with known variance and a uniform prior

- **Model**

$$y_1, ..., y_n | \theta \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2), \quad [\sigma^2 \text{ is } \textbf{known}.]$$

- **Prior**

$$p(\theta) \propto c \text{ (a constant)}.$$

- **Likelihood** [**white board!**]

$$\begin{aligned} p(y_1, ..., y_n | \theta) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{y})^2\right]. \end{aligned}$$

- **Posterior**

$$\theta | y_1, ..., y_n \sim \mathcal{N}(\underbrace{\bar{y}}_{\mu_n}, \underbrace{\sigma^2/n}_{\tau_n^2})$$

- **Make your life easy**: **throw away** normalizing constants independent of θ!

Some remarks

- **The prior** $p(\theta) \propto c$ is **improper**: $\int p(\theta)d\theta = \infty$.
- **WARNING: improper priors** may lead to **improper posteriors**. **Not valid** since the posterior is (by construction) a probability distribution

$$\int p(\theta|y)d\theta = 1.$$

- This prior is said to be **non-informative** because it **does not favor** any θ a priori. More on this later.
- The **posterior mode** is $\mu_n = \bar{y}$. Coincides with the **maximum likelihood estimator**

$$\mu_{\mathrm{MLE}} = \bar{y}.$$

- We will learn that **Bayesian inference with a non-informative prior** gives the same **point estimates** as classical inference...
- ... but a **different interpretation**

Posterior = a probability distribution = **fun inference!**

- **Prior**

$$\theta \sim \mathcal{N}(\mu_0, \tau_0^2).$$

- **Posterior**

$$\begin{aligned} p(\theta|y_1, ..., y_n) &\propto p(y_1, ..., y_n|\theta)p(\theta) \\ &\propto \mathcal{N}(\theta|\mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \quad \textbf{and} \quad \mu_n = w\bar{y} + (1-w)\mu_0$$

with

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

- **White board**: **tedious** but a good exercise. **Hint:** the joy of ignoring a constant.
- **Interpretation** of the posterior as a **combination of prior and data information**.

A combination of prior and data information

- Define the **precision** as the **reciprocal** of the variance

$$\text{Precision} = \text{Variance}^{-1} = \frac{1}{\text{Variance}}$$

- Thus

$$\text{Prior Precision} = \frac{1}{\tau_0^2}, \quad \text{Data Precision} = \left(\frac{\sigma^2}{n}\right)^{-1} = \frac{n}{\sigma^2}$$

- Reading the equations **out loud** (for the normal model with normal prior)

$$\begin{aligned}\text{Posterior Precision} &= \text{Data Precision} + \text{Prior Precision}\\ \text{Posterior mean} &= \underbrace{\frac{\text{Data Precision}}{\text{Posterior Precision}}}_{w} \times (\text{Data mean})\\ &+ \underbrace{\frac{\text{Prior Precision}}{\text{Posterior Precision}}}_{1-w} \times (\text{Prior mean}).\end{aligned}$$

Some remarks

- Note that, informally, if $\tau_0^2 = \infty$ then $\text{Prior Precision} = 0$ and

$$\text{Posterior mean} \quad = \quad \text{Data mean}$$

- The **improper prior** $p(\theta) \propto c$ is the limit

$$p(\theta) = \mathcal{N}(\mu_0, \tau_0^2), \quad \text{when } \tau_0^2 \to \infty.$$

- **My two cents**: I never use "non-informative" priors. I choose a proper prior with hyper parameters that allow **for a wide range** of θ values instead...
- ... but **only if I lack prior** information. Otherwise I incorporate prior information.
- **Don't be ashame** of using priors.
  - It is **a part of your model**.
  - Is someone **accusing you** for a subjective analysis? **My two cents**: nothing in a model is more subjective than the model itself.
  - It **makes sense to include prior information**. E.g. if doing a study on a drug, previous experiments are interesting.
  - Prior information is **often implicitly used in classical statistics**. But it is hidden!