ST308 - Lent term
Bayesian Inference

Kostas Kalogeropoulos

Introduction to Bayesian Inference

Outline

Outline

Teaching

My name is Kostas Kalogeropoulos and will be doing the lectures.

Phil Chan and Patrick Aschermayr will be doing the computer classes.

Lectures: 2 hours every week on Friday 10:00–12:00 in room CLM.3.02 .

Computer Classes: Thursdays 13:00-14:00 (PC) and 14:00-15:00 (PC), Fridays 13:00-14:00 (PA).

In Class

During lectures we will cover the theory and go through several examples.

Recordings will be available on Moodle but try to attend lectures anyway.

During computer classes we will go through the computer part of the course.

Class attendance is compulsory and will be recorded on LSE for You.

Moodle

All relevant material will be posted on Moodle.
Enrolment key: Laplace

- Slides for each lecture.
- Recording of lectures.
- R Markdown notebooks for each computer class
- Problem sets each Friday.
- Solutions of problem sets with code once you have handed them back.

Computing

- RStudio will be used throughout the course.
- You can either bring your laptop to the computer classes or use the room's PC.
- Install the latest R and RStudio version.

Weekly formative assignments

Each week you will be assigned a problem set containing both theoretical and computer exercises.

It will be due on your computer class day of next week. Submitted in the computer class.

Hand in everything you are able to solve even if it is not complete. Marks don't count in the final grade but are recorded on LSE for You.

Problem sets will be returned marked with feedback during the class of the week after you handed them in.

Assessment

An individual project will be assigned on week 7 and will be due Tuesday, May 12th noon. You will be required to analyse data of your choice using the taught Bayesian Inference techniques and present your findings through a paper-like report.

During summer term the course is assessed by a 2 hour written exam.

The final grade will be determined by the above with weights (20-80%).

Questions and Feedback

1. Ask questions in class.
2. Office and Feedback hours: COL.6.10 on Monday 10:30–12:30.
3. Use the forum on Moodle.
4. Feel free to send emails but please try to avoid questions that might be of interest for all the class; use Moodle forum instead.

Syllabus and plan

- Week 1: Introduction to Bayesian Inference
- Week 2: Statistical Decision Theory - Bayes Estimators
- Week 3: Credible Intervals - Priors - Multiparameter models
- Week 4: Bayesian Model Choice and Prediction - Monte Carlo
- Week 5: Bayesian linear regression
- Week 6: Bayesian Classification
- Week 7: Markov Chain Monte Carlo sampling - Examples of Bayesian inference
- Week 8: Variational Bayes
- Week 9: Mixture models and the EM algorithm
- Week 10: Gaussian Processes Regression and Classification

Readings

Lecture slides are sufficient for exam purposes.

For further optional reading check

- J.O. Berger, *Statistical Decision Theory and Bayesian Analysis.*
- D. Gamerman, H. F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.*

Other examples are

- J.K. Kruschke, *Doing Bayesian Data Analysis. An tutorial with R, JAGS and Stan.* 2nd edition.
- Gelman, Carlin, Stern and Rubin. *Bayesian Data Analysis*,
- C.P. Robert, *The Bayesian Choice*

Books (cont'd)

Also, for non-academic textbooks you can check the following

- Nate Silver *The Signal And The Noise. Why Most Predictions Fail – but Some Don't.*
  - Named Amazon's #1 Best NonFiction Book for 2012.
  - Reached #4 in New York Times Best Sellers list for non-fiction books.
- Sharon Bertsch McGrayne *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*

Outline

Quotes about Bayesian Inference (positive)

‘Every statistician would be a Bayesian if he took the trouble to read the literature thoroughly...’ (D. V. Lindley)

‘... everyone is, should be, or will soon be a Bayesian’ (H. Chernoff)

‘... the best way to convey to the experimenter what the data tell him about is to show him a picture of the posterior distribution.’
(G. E. P. Box & G. C. Tiao)

Quotes about Bayesian Inference (negative)

'...**Bayes,** who seems to have first attempted to apply the notion of probability to causes in relation to their effects, **invented a theory, and evidently doubted its soundness, for he did not publish it during his life.** It was posthumously published by Price, who seems to have felt no doubt of its soundness.'
(R. A. Fisher 1930)

[Referring to Bayesians]: 'If they would only do as he (Bayes) did and publish posthumously we should all be saved a lot of trouble.'
(M. G. Kendall)

Historical remarks

- The term Bayesian refers to reverend Thomas Bayes, who proved a special case of the Bayes theorem.
- Pierre-Simon Laplace introduced a general version of the theorem and used it to approach problems in celestial mechanics, medical statistics, reliability, etc
- Initially it was called inverse probability as it inverts from observations to parameters, or from effects to causes.
- Until 1980 frequentist (aka classical) inference was used by the majority of statisticians.

Historical remarks (cont'd)

- The development of Markov Chain Monte Carlo algorithms (MCMC) removed many of the computational problems.
- Today Bayesian inference implemented via MCMC may be used to nonstandard and highly complex applications where often the frequentist approach is infeasible.
- A list of areas of applications includes artificial intelligence, biostatistics, econometrics, epidemiology, finance, genomics, geostatistics, image processing and pattern recognition, neural networks, signal processing etc.

Defining probability

Frequentist definition: If the experiment was repeated many times, probability of A is the frequency f_n in which a given event A is realised.

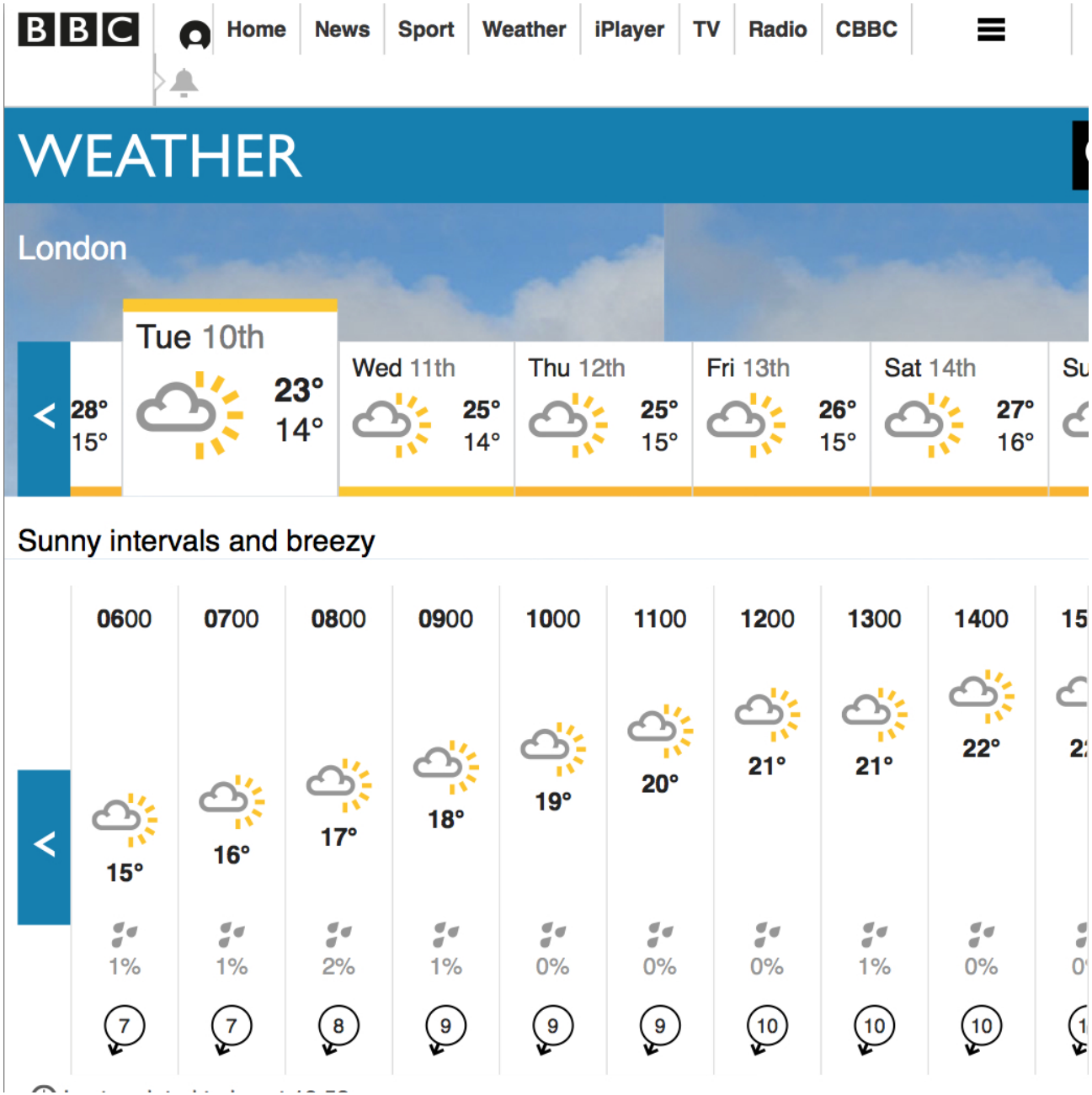
$$P(A) = \lim_{n \to \infty} f_n.$$

Subjective definition: $P(A)$: a number in $[0, 1]$ reflecting our beliefs on how likely A is (0:impossible, 1:certain).

Example: Probability of heads when tossing a fair coin?

- If we toss a fair coin n times, where n is large, count the number of times we get head say k_n over n. If the coin is fair, we should get 0.5.
- Alternatively, we can argue that it makes sense for this probability to be 0.5 as we believe that heads and tails are equally likely.

Probability of rain in a particular day?



Frequentist Probability and Time Travel?

Frequentist Probability and Multiverse?

Data and likelihood

- Consider data $y = (y_1, y_2 \ldots, y_n)$ from a real world application.
- Assign a suitable probability model aka likelihood for data y and parameter(s) $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

$$f(y_1, \ldots, y_n | \theta_1, \ldots, \theta_p) = f(y|\theta).$$

- Use the above to learn about θ and answer the relevant questions or predict future y.

Prior Information

Consider the following 3 experiments where the probability of a correct answer θ is of interest.

1. A lady claims that by tasting a cup of tea with milk she can tell whether the milk was poured into the cup before the tea. In 9 out of 10 trials she gets it right.
2. A musical expert claims that he can distinguish by a small music part whether it is Mozart or Haydn. He gets it right in 9 out of 10 times.
3. A drunk man claims he can predict the outcome of a fair coin flip. In 9 flips out of 10 he is correct.

Frequentist approach: Test $H_0 : \theta = 0.5$ vs $H_1 : \theta > 0.5$.
This gives p-value ≈ 0.01 concluding genuine skill in all cases.

Bayes theorem for events

Clearly there exists some relevant information prior to the experiments. Bayesian inference formally seeks to utilise prior information on θ by the so-called prior distribution $\pi(\theta)$

Bayes Theorem for Events: In terms of events and their probabilities, let A and B be two events such that $P(A) > 0$. then $P(B|A)$ and $P(A|B)$ are related by

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$$

More generally if $B_1, B_2, \ldots, B_k$ form a partition (k can be ∞), we can write for all $j = 1, \ldots, k$

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}, \text{ where } P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Bayesian Statistical model

We treat θ as a random variable and assign the prior pdf $\pi(\theta)$. The prior reflects our beliefs on θ before seeing the data.

The posterior distribution reflects our beliefs on θ after seeing the data and is the main object for inference. It is given by

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}, \text{ where } f(y) = \int f(y|\theta)\pi(\theta)d\theta$$

The term $f(y)$ is known as marginal likelihood or evidence and reflects probability of the data under the adopted probability model. It may also be viewed as a normalising constant.

Features of Bayesian Inference

- **Prior information:** Every problem is different and has its own context which can (in theory) be reflected via the prior distribution.

  Criticism: It is not always clear how to define the prior distribution. Results depend on the choice of it.

- **Subjective probability:** Accepting the subjective basis of knowledge. Also always defined contrary to the frequentist definition that applies to inherently repeatable events.

  Criticism: No guarantee that your quantification of uncertainty will be seen by others as 'good'.

- **Common misconception:** Bayesians believe that there is a single unknown value for θ. A distribution is assigned only to express subjective uncertainty not because the truth is random.

Steps of Bayesian Inference

1. **Likelihood:** Same as in frequentist approach. Sometimes indicated by the data observation mechanism, sometimes assumptions are made and checked afterwards.

2. **Prior:** See material of week 3.

3. **Calculation to get the posterior:** See what follows.

4. **Inference from the posterior:** See material of weeks 2, 3 and 4.

Outline

Example 1: Binomial-Beta

Suppose that x is a single observation from a Binomial(n, θ) random variable.

1. **Likelihood:** The likelihood is given by the probability of x given θ which is provided by the Binomial distribution

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \propto \theta^x (1-\theta)^{n-x}$$

2. **Prior:** As $0 < \theta < 1$ a corresponding distribution must be chosen. The Beta distribution with hyper-parameters α and β, denote Beta(α, β), provides such an example.

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Example 1: Binomial-Beta (cont'd)

3. **Posterior:**

$$\begin{aligned}\pi(\theta|x) \propto f(x|\theta)\pi(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1} \\ &\overset{\mathcal{D}}{=} \text{Beta}(\alpha+x, n-x+\beta)\end{aligned}$$

Note: The posterior mean, that could be used as an estimator for p is equal to $\frac{\alpha+x}{\alpha+\beta+n}$. For $\alpha = \beta = 0$ it coincides with the MLE x/n. No such prior distribution (improper prior) but posterior is well defined.

A trick for finding posterior distributions

Since $m(x) = \int f(x|\theta)\pi(\theta)d\theta$ is independent of θ, we can write $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

- If $\theta|x \sim N(\mu, \sigma^2)$, for σ^2 known, then

$$p(\theta|x) \propto \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{\theta^2 - 2\theta\mu}{2\sigma^2}\right)$$

- If $\theta|x \sim \text{Gamma}(\alpha, \beta) \rightarrow \pi(\theta|x) \propto e^{-\beta\theta}\theta^{\alpha-1}$
- If $\theta|x \sim \text{Beta}(\alpha, \beta) \rightarrow \pi(\theta|x) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$.

Work out $f(x|\theta)\pi(\theta)$ in terms of θ and inspect the above (or other known distributions).

Example 2: Poisson-Gamma

Let $x = (x_1, \ldots, x_n)$ be a random sample (x_i's are independent and identically distributed) from a Poisson(λ) population.

1. **Likelihood:** The likelihood is given by the joint density of the sample

$$f(x|\lambda) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \propto \exp(-n\lambda)\lambda^{\sum x_i}$$

2. **Prior:** As $\lambda > 0$ a corresponding distribution must be chosen. The Gamma distribution with hyper-parameters α and β, denote Gamma(α, β), provides such an example.

$$\pi(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

Example 2: Poisson-Gamma (cont'd)

3. **Posterior:**

$$\begin{aligned}\pi(\lambda|x) \propto f(x|\lambda)\pi(\lambda) &\propto \exp(-n\lambda)\lambda^{\sum x_i}\lambda^{\alpha-1}\exp(-\beta\lambda) \\ &= \lambda^{\alpha+\sum x_i-1}\exp(-n\lambda-\beta\lambda) \\ &= \lambda^{\alpha+\sum x_i-1}\exp\left[-\lambda(n+\beta)\right] \\ &\overset{\mathcal{D}}{=} \text{Gamma}(\alpha+\sum x_i, n+\beta)\end{aligned}$$

Note: The posterior mean can be written as

$$\frac{\alpha+n\bar{x}}{n+\beta} = \ldots = \left(1-\frac{n}{n+\beta}\right)\frac{\alpha}{\beta}+\frac{n}{n+\beta}\bar{x},$$

which is a weighted average between the prior mean and $\bar{x}$. As $n \to \infty$ the posterior mean converges to $\bar{x}$.

Example 3: Normal-Normal

Let $x = (x_1, \ldots, x_n)$ be a random sample (x_i's are independent and identically distributed) from the N(θ, σ^2) - σ^2 known.

1. **Likelihood:** The likelihood is given by the joint density of the sample

$$f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}\right)$$

2. **Prior:** We assume another Normal prior for θ, N(μ, τ^2), which gives

$$\pi(\theta) \propto \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right)$$

Example 3: Normal-Normal (cont'd)

3. **Posterior:**

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}\right)\exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right)$$

$$\propto \ldots^{1} \propto \exp\left(-\frac{\theta^2 - 2\theta\frac{\frac{\sigma^2}{n}\mu+\tau^2\bar{x}}{\tau^2+\frac{\sigma^2}{n}}}{2\frac{\tau^2\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}}\right) \overset{\mathcal{D}}{=} \mathrm{N}\left(\frac{\frac{\sigma^2}{n}\mu+\tau^2\bar{x}}{\tau^2+\frac{\sigma^2}{n}}, \frac{\tau^2\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\right)$$

Notes on Normal-Normal model

- The posterior mean can be written as a weighted average between the prior mean and $\bar{x}$

$$\left(1 - \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\right)\mu + \frac{\tau^2}{\frac{\sigma^2}{n} + \tau^2}\bar{x}$$

- Fisher information equals inverse variance, aka precision. We can write the posterior precision as the sum of the prior precision and $n\times$ the precision of each observation

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

Notes on Normal-Normal model (cont'd)

- As $n \to \infty$ the posterior distribution converges to

$$\mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

  which is independent of the prior.

- As $\tau^2 \to \infty$ the posterior distribution converges again to

$$\mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

Reading

J.O. Berger Sections 1.1, 1.2, 4.1 and 4.2

D. Gamerman & H.F. Lopes 2.1 2.2 2.3