Elementary Bayesian Analysis

- Credits: 9 cr
- Teacher: Doc. Dr.Tech Aki Vehtari
- Contents: Bayesian probability theory and Bayesian inference. Bayesian models and their analysis. Computational methods, Markov chain Monte Carlo.
- Requirements: Exam and exercises.
- Literature: Gelman, Carlin, Stern & Rubin: Bayesian Data Analysis, Second edition, and other material announced on course web page.
- Prerequisites: Basics in statistical mathematics and probability calculus.

Slide 1

Literature

- Gelman, Carlin, Stern & Rubin: Bayesian Data Analysis, Second edition
  - the chapters 1-16 and 22
  - excellent book, consider buying it
- Other material announced on the course web page
  - available online

Slide 2

Contents (based on Gelman et al)

Slide 3

- Introduction (ch 1)
- Single- and multiparameter-models (chs 2,3)
- Large-sample inference and frequency properties of Bayesian inference (ch 4)
- Hierarchical models (ch 5)
- Computational methods, Markov chain Monte Carlo (chs 10-13)
- Decision analysis (ch 22)
- Model checking and improvement (ch 6)
- Modeling accounting for data collection (ch 7)
- Connections, challenges, general advice (chs 8-9)
- Regression (chs 14-16)

Requirements

Slide 4

- Exam
  - exam questions based on the book only
  - few example exams will be available on the course web page
- Exercises
  - listed on the course web page
  - some pen and paper exercises
  - some computer exercises (with Matlab)
  - group work allowed (1–3 students)

Computer exercises

- Computer exercises have been picked mostly from the book
  - there are hints and templates for doing the computer exercises with Matlab
  - there will be several Matlab demos on the course web page, illustrating some of the concepts and algorithms

Slide 5

Introduction

- Some application areas
- History of term "Bayesian"
- Probability as a measure of uncertainty
- Combining uncertain information using probability calculus

Slide 6

- Bayes' rule
- Bayesian model
- About computation

Slide 7

Some application areas

- Archeology
- Astronomy
- Bio-sciences
- Cognitive science
- Data mining
- Decision analysis
- Economy
- Epidemiology
- Genetics
- Image analysis
- Law
- Medicine
- Meteorology
- Physics
- Process modeling
- Reliability analysis
- Signal analysis
- Social sciences
- *Anything related to real world, where inference is made based on observations*

Slide 8

Some projects in BECS

- MEG brain imaging
- ECG/MCG based disease risk prediction
- Spatial epidemiology
- Modeling of healthcare processes
- Bio-spectroscopy
- Modeling and prediction of concrete quality
- Modeling and prediction of continuous steel casting
- Electrical impedance tomography of industrial pipes
- Modeling of human perception
- Machine vision

Bayesian Analysis

Slide 9

- Based on Bayesian probability theory
  - uncertainty is presented with probabilities
  - probabilities are updated based on new information
  - *...common sense reduced to calculation*, Laplace 1819
- Thomas Bayes (170?–1761)
  - English nonconformist, Presbyterian minister, mathematician
  - Richard Price published Bayes' paper on conditional probabilities in 1763 after Bayes had died
  - considered the problem of *inverse probability*
    - significant part of the Bayesian theory
- Bayes did not invent all, but was first to solve problem of inverse probability in special case
- Modern Bayesian theory with rigorous proofs developed in 20th century

Term Bayesian used first time in mid 20th century

Slide 10

- Earlier there was just "probability theory"
  - concept of the probability was not strictly defined, although it was close to modern Bayesian interpretation
  - in the end of 19th century there were increasing demand for more strict definition of probability (mathematical and philosophical problem)
- In the beginning of 20th century frequentist view gained popularity
  - accepts definition of probabilities only through frequencies
  - does not accept inverse probability or use of prior
  - gained popularity due to apparent objectivity and "cook book" like reference books
- Frequentist R. A. Fisher used in 1950 first time term "Bayesian" to emphasize the difference to general term "probability theory"
  - term became quickly popular, because alternative descriptions were longer
  - after this Bayesians started to use term "frequentist"

Popularity of Bayesian methods increasing

- Modern Bayesian theory based on axioms, developed in 20th-century, solved the mathematical problem
  - philosophical argument with frequentist continued
- Increased computing speed has allowed use of full power of Bayesian approach in modeling more complex phenomena, making it's popularity to soar
  - most users of the methods are pragmatic, that is, they use the methods because they work well

Slide 11

Bayesian analysis

- Uncertainty is described with probability
- Uncertainties are combined using probability calculus

Slide 12

Probability as a measure of uncertainty

Slide 13

- A event, I background information
- $p(A|I)$ probability of A given I
  Measures uncertainty based on information I:
  - $p(A|I) = 1$ if you are sure, that A happens
  - $p(A|I) = 0$ if you are sure, that A does not happen
  - $p(A|I) = 0.4$: there is uncertainty whether A happens (but *not* necessarily randomness)
  - if it is more likely that A happens compares to B then $p(A|I) > p(B|I)$

Aleatory vs. epistemic uncertainty

Slide 14

Uncertainty can be divided in

- Aleatory uncertainty, due to randomness
  - we are not able to get observations, which would help to reduce this uncertainty
- Epistemic uncertainty, due to lack of knowledge
  - we can get observations, which help to reduce this uncertainty
- Coin example
  - two observers may have different epistemic uncertainty
  - epistemic uncertainty changes, when information changes

Example: Colored chips in a bag

- If proportion of the different colors is known
  - aleatory uncertainty about what color will be picked next
- If proportion of the different colors is known
  - additional epistemic uncertainty
  - epistemic uncertainty changes when colors of chips are observed

Slide 15

- If instead of picking chips one at time, we empty the whole bag and count the proportion of the colors
  - no aleatory uncertainty
  - only epistemic uncertainty about the contents of the bag

Combining uncertainties?

- Let's mark
  - y observed chips
  - θ proportion of the colors
  - I background knowledge
- Aleatory uncertainty, if proportion of chips θ is known

Slide 16

$$p(y|\theta, I)$$

- Epistemic uncertainty before observations

$$p(\theta|I)$$

- How to update epistemic uncertainty when we observe chips?

$$p(\theta|y, I)?$$

  - θ is unknown, i.e., we want to know the inverse probability

Axiomatic justifications for using probabilities*

- Probability as a measure of uncertainty and probability calculus can be justified axiomatically
  - some variations, with same basic ideas, but slightly different presentations
  - two basic approaches
    - probability and utility separated (e.g.. Cox, DeGroot, ...)
    - probability and utility non-separable (e.g.. de Finetti, Savage, Bernardo & Smith, ...)

Slide 17

A axiomatic formulation (in words)*

(A1) All events can be compared

(A2) Comparisons are transitive
and arbitrary accurate comparisons can be made

(A3) No event is more unlikely than certainly false
and certainly false is more unlikely than certainly true

Slide 18

(A4) Finite additivity

(A5) Quantification (e.g. idealized roulette)
- unique quantitative probability values

Axiomatic formulation*

Probability calculus is derived from axioms, and that's all you need

(P1) $p(\text{event}) \geq 0$ and $p(\text{certainly true}) = 1$.

(P2) Sum-rule

if A and B exclusive, then $p(A, B) = p(A) + p(B)$

Slide 19

(P3) Sum-rule for infinite sequences

(P4) Bayes' rule

$p(A|B) = p(A, B)/p(B)$

- product-rule can be derived from this

  $p(A, B) = p(B|A)p(A)$

If uncertainties are combined in some other way, then at least one of the axioms is not respected

Bayes' rule

- We can select $p(y|\theta, I)$ and $p(\theta|I)$, and then compute using Bayes' rule

$$p(\theta|y, I) = \frac{p(y|\theta, I)p(\theta|I)}{p(y|I)}$$

Slide 20

- Names for different parts
  - $p(\theta|y, I)$ = posterior
  - $p(y|\theta, I)$ = model or likelihood
  - $p(\theta|I)$ = prior
  - $p(y|I)$ = normalization (evidence)

Model / likelihood

- Model $p(y|\theta, I)$
  - mathematical description of the observation model / data generating process / aleatory part
  - if phenomenon is known with given θ, I what is the probability that we would observe y with given value
- Likelihood $p(y|\theta, I)$

Slide 21

  - when regarded as a function of θ, for fixed y, it is called the *likelihood function*
  - note that likelihood term is often misused as a replacement for term model

Bayesian model parts

- Prior $p(\theta|I)$
  - mathematical description what is known about θ
  - epistemic uncertainty before observations
  - model (likelihood) and prior inseparable

+ if phenomenon is known, no epistemic uncertainty

Slide 22

+ if no observations, epistemic uncertainty does not change

Bayes' rule

- Posterior distribution describes the updated epistemic uncertainty when information from the observations and the prior is combined

$$\begin{aligned} p(\theta|y, I) &= \frac{p(y|\theta, I)p(\theta|I)}{p(y|I)} \\ &= \frac{p(y|\theta, I)p(\theta|I)}{\int p(y|\theta, I)p(\theta|I)d\theta} \\ &\propto p(y|\theta, I)p(\theta|I) \end{aligned}$$

Slide 23

- Normalization term normalizes the total probability in the posterior to 1

Where do we get I, $p(\theta|I)$, and $p(y|\theta, I)$?

- Excellent question!
- Same problem in non-Bayesian approaches!
  - model has to be chosen
  - often there is also something, like regularization term, which corresponds to prior $p(\theta|I)$
  - leaving out prior term corresponds to to uniform prior on θ

Slide 24

Subjectivity

- Aleatory uncertainty apparently objective
  - choosing model which describes aleatory uncertainty is subjective
- Epistemic uncertainty is clearly subjective
  - conditioned on the knowledge of the observer
  - different observers may have different opinion about uncertainty (“different I”)

Slide 25

- Scientific objectivity can be achieved by inter-subjectivity
  - if scientist agree on assumptions made (“same I”)

Bayesian analysis

- Model
  - tries to model behavior of a phenomenon
  - often simplifies the reality
  - may be used to predict future
  - may be used to gain scientific insight

Slide 26

- Simplifies because
  - observations are not direct or have limited accuracy
  - some observables may have much larger effect than others
  - simplified model may still be useful for practical purposes

Example

- Ball is dropped from different heights, and time is measured using hand hold stop watch
  - Newton’s mechanics
  - air resistance, air pressure, shape of the ball, surface structure of the ball
  - air currents
  - relativity

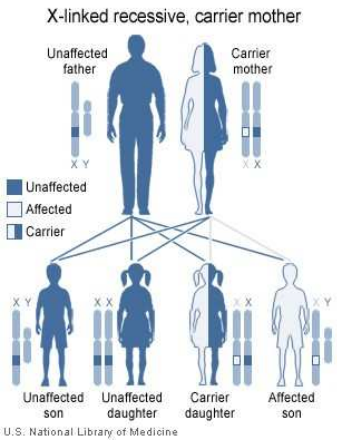
Slide 27

- Given observations, how accurate model it is worth making?
- There are many situations where, simpler models are useful and practically as accurate as more complex models!
- "*All models are wrong but some models are useful*", George P. Box

Example: Hemophilia

- Hereditary disease, X-chromosome-linked recessive
- A woman has an affected brother, her mother and father are healthy

Slide 28



Example: Hemophilia

- Hereditary disease, X-chromosome-linked recessive
- A woman has an affected brother, her mother and father are healthy
- Based on background knowledge let's form a model M
  - model is simplified from the reality, since
    - possibility of twins is ignored
    - possibility of mutation is ignored
    - etc.

Slide 29

Example: Hemophilia

- Hereditary disease, X-chromosome-linked recessive
- A woman has an affected brother, her mother and father are healthy
- She is carrier ($\theta = 1$) or not ($\theta = 0$)
  $p(\theta = 1|M) = p(\theta = 0|M) = \frac{1}{2}$
- She has 2 healthy sons
  $p(y_1 = 0, y_2 = 0|\theta = 1, M) = (0.5)(0.5) = 0.25$
  $p(y_1 = 0, y_2 = 0|\theta = 0, M) = (1)(1) = 1$
- Posterior
  $$p(\theta = 1|y, M) = \frac{p(y|\theta = 1)p(\theta = 1)}{p(y|\theta = 1)p(\theta = 1) + p(y|\theta = 0)p(\theta = 0)}$$
  $$p(\theta = 1|y, M) = \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.2$$

Slide 30

Prediction

- E.g., $y = (y_1, \ldots, y_n)$ are observations
- $\tilde{y}$ is new observations which has not yet been made
  - $\tilde{y}$ is unknown and thus there is uncertainty
- prediction for $\tilde{y}$

Slide 31

$$\begin{aligned} p(\tilde{y}|y, M) &= \sum_{\theta=0,1} p(\tilde{y}|\theta, y, M)p(\theta|y, M) \\ &= \sum_{\theta=0,1} p(\tilde{y}|\theta, M)p(\theta|y, M) \end{aligned}$$

- Uncertainty in prediction includes both aleatory and epistemic uncertainty

Example: Hemophilia

- Third son?
  $p(y_3 = 0|y_1, y_2, M)$
- Prediction

Slide 32

$$\begin{aligned} p(y_3 = 0|y_1, y_2, M) &= \sum_{\theta=0,1} p(y_3 = 0|\theta, M)p(\theta|y_1, y_2, M) \\ p(y_3 = 0|y_1, y_2, M) &= p(y_3 = 0|\theta = 1, M)p(\theta = 1|y_1, y_2, M) \\ &\quad + p(y_3 = 0|\theta = 0, M)p(\theta = 0|y_1, y_2, M) \\ p(y_3 = 0|y_1, y_2, M) &= (0.5)(0.2) + (1)(0.8) = 0.9 \end{aligned}$$

Example: Hemophilia

- Third son is healthy
  - new observations can be used to update uncertainty about her state
- Chain rule
  - previous posterior is now new prior

Slide 33

$$\begin{aligned} p(\theta = 1|y_1, y_2, y_3) &= \frac{p(y_3|\theta = 1, M)p(\theta = 1|\mathbf{y}_1, y_2, M)}{\sum_{\theta=0,1} p(y_3|\theta, M)p(\theta|\mathbf{y}_1, y_2, M)} \\ &= \frac{(0.5)(0.2)}{(0.5)(0.2) + (1)(0.8)} = 0.111 \end{aligned}$$

Integration in Bayesian approach

- Summing generalizes to integration for continuous variables
  - often notation is simplified and integration symbol is used for discrete variables, too
- Normalization

$$p(y|M) = \int p(y|\theta, M)p(\theta|M)d\theta$$

Slide 34

- Prediction

$$p(\tilde{y}|y, M) = \int p(\tilde{y}|\theta, M)p(\theta|y, M)d\theta$$

- Marginalization

$$p(y|\theta_1, M) = \int p(y|\theta_1, \theta_2, M)p(\theta_2|M)d\theta_2$$

Integration in Bayesian approach

- Integration replaced by optimization: maximum a posterior (MAP)
  - works in simple cases
- Analytical integration
  - works with some simple models
- Analytic approximations

Slide 35

  - work with simpler or restricted models, and may need lot of work
- Numerical integration
  - needs computational power

Numerical Integration

- *Monte Carlo* (MC)
  - integral is approximated using samples drawn from the posterior distribution ($A^{(t)}$)

$$E(A) \approx \frac{1}{N} \sum_{t=1}^{N} A^{(t)}$$

  - often difficult to get independent samples efficiently

Slide 36

- *Markov Chain Monte Carlo* (MCMC)
  - uses Markov chains
  - dependent samples (makes accuracy estimation more difficult)
  - popularity soared in 1990's

Increase in popularity of Bayesian methods

- mainly analytic approximations were used until 1990's
  $\longrightarrow$ models had to be simple
- Increase in computational power and development of the MCMC methods
  $\longrightarrow$ popularity soared in 1990's
  $\longrightarrow$ possibility to use more complex models describing reality better
  $\longrightarrow$ methods were taken in use in many areas having difficult problems

Slide 37

Summary of Bayesian analysis

- Based on background information form a model
  - model / likelihood
  - structural prior
- Supplement with possible background information
  - prior for parameters
- Use Bayes' rule and marginalization to compute probability distributions for desired unknowns
  - e.g. prediction for future observation

Slide 38

Example: Mass of Saturn

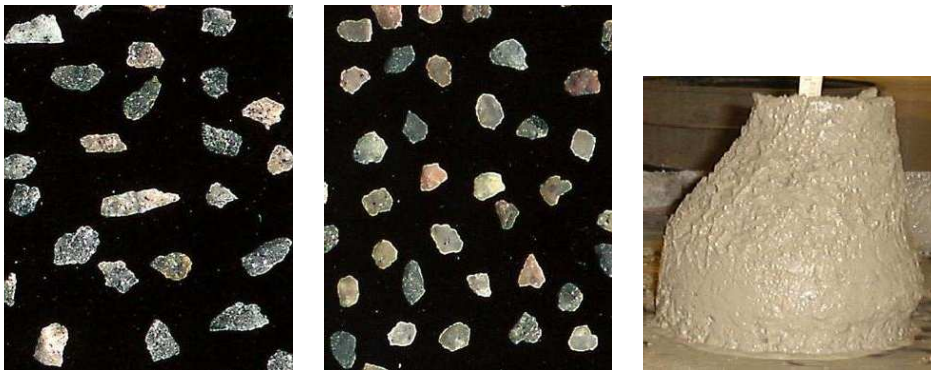
Slide 39

- Model and observations
  - θ = mass of Saturn (unknown)
  - D = disturbances in the orbits of the Jupiter and Saturn measured by observatories (observations)
  - M = Newton's mechanics (modeling assumptions)
  - $p(D|\theta, M)$ = if mass of the Saturn were θ, how likely it would be to observe D (model/likelihood)
  - $p(\theta|M)$ = sensible restriction to mass; not too small to Saturn lose its rings, not too large to break whole solar system (prior)
- Laplace calculated and stated '.....it is a bet of 11,000 to 1 that the error in this result is not 1% of its value'
  - the modern estimate differs from Laplace's by 0.63%
- Note, that Laplace calculated also the uncertainty of the estimate, and thus was able give accuracy estimate

Example: Prediction of concrete quality

Slide 40

- Effect of recipe and aggregates to concrete quality
  - amount of water, cement, aggregates, and additives
  - physical and chemical properties of aggregates



Example: Prediction of concrete quality

- Gaussian process model
  - non-linear regression model
  - similar recipe and stone material produce similar quality
  - similarity is described wit covariance function, which parameters are unknown
- Using the model and conclusions made by Dr.Tech Hanna Järvenpää after examining the phenomena using the model it was possible
  - to reduce material costs by 5-15%
  - to reduce proportion of natural gravel in concrete to 5-20% compared to previous 60-100%

Slide 41