Statistical Methods
(summer term 2024)

Bayesian parameter estimation

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Overview

- Parameter estimation from the Bayesian point of view

- Related to maximum likelihood estimation (MLE)

- Mostly concerned with how to incorporate prior information to MLE, and paying more attention to the resulting PDF
  - MLE provides so-called $point\ estimates$, while Bayesian estimation provides probability distributions for parameters

- As the title suggests, it is a parametric method (i.e. it makes assumptions about the underlying distribution of the data)
  - aim: obtaining the full (joint) PDF of the probability (density/mass) function of parameters
  - allows one to derive summary statistics for parameters, such as means, variances, covariances, etc.

Bayesian parameter estimation – basic idea

- MLE: set-up the likelihood function and study its properties around the maximum

- Bayesian: use Bayes' theorem to obtain the probability of parameters θ given the data D

$$P(\theta|D,M) = \frac{P(D|\theta,M)\,P(\theta,M)}{P(D|M)}$$

- M is a model representing the background information, e.g. that the data were drawn from a Normal distribution $\rightarrow$ *generative model*

Bayesian parameter estimation – basic idea

- The Bayesian interpretation of this relation for the model parameters θ is

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- This relation is of deceiving apparent simplicity
  - θ can be a parameter vector $\rightarrow$ potentially high dimensional problem
  - the evidence (probability of the data) is often difficult to calculate
  - not always possible to express prior information as a well-defined probability

Bayesian parameter estimation – dealing with the problems

$$P(\theta|D,M) = \frac{P(D|\theta,M)\,P(\theta|M)}{P(D|M)} \qquad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- The evidence (probability of the data) is independent of the parameters
  - it is a normalization constant that can be ignored as long as one is only interested in the shape of the posterior
    - ⋆ the location of the maximum of the posterior is not affected
  - once the posterior is obtained, it can be (perhaps numerically) normalized

- Improper priors: sometimes the prior information, another model, or common sense cannot be normalized
  - e.g., when measuring a length s, it should be a positive quantity $s > 0$
    How would you formulate probability density on s?

- A frequently used shorthand notation is this

$$P(\theta|D,M) \propto L(D|\theta,M)\,P^*(\theta|M) \qquad \text{posterior} \propto \text{likelihood} \times \text{prior}$$

the "$*$" indicates a possibly improper or non-normalized prior

The evidence as marginal probability

- We mentioned that $P(D|M)$ is called the evidence of model M
  - We will use this later for *model selection*
  - It allows us to compare different models to decide which one is more likely to be correct given the data

- Using the law of total probability, the evidence can be written as:

$$P(D|M) = \int P(D|\theta, M)\, P(\theta|M)\, d\theta = \int \text{likelihood}(D|\theta, M) \times \text{prior}(\theta|M)\, d\theta$$

- The evidence measures how well the model predicts the data, independent of the specific parameter values θ
  - It can be thought of as an average predictive ability of the model over all plausible parameter values
  - for a "fair" evaluation of the evidence, the prior must be normalized with respect to the parameters θ, and the likelihood with respect to the data D, ensuring it is a proper PDF with respect to the data

The evidence as marginal probability

- The product of likelihood and prior can be also seen as the *joint* probability density of data and parameters

$$P(D \cap \theta|M) = P(D|\theta, M)\, P(\theta|M)$$

- The process of integrating over one parameter is called marginalization
  - Statistics lingo: We "marginalize out" a parameter
  - One can also marginalize out more parameters to obtain the probability distribution of the remaining parameters

- As the heading suggests, the evidence can be seen as the marginal probability of the joint probability of data and parameters
  - This means we marginalize out the parameter θ to obtain the probability of the data given the model

$$P(D|M) = \int P(D|\theta, M) P(\theta|M) d\theta$$

Result of Bayesian estimation: (joint) PDF of the parameter(s)

- The result of Bayesian estimation is the full probability density/mass function of the parameters
  - In multidimensional problems, this results in the joint PDF of the parameters

- It provides a complete statistical description of the parameters

- Often, we wish to summarize/characterize the posterior PDF in terms of:
  - MAP (maximum a posteriori) estimate = mode, mean, or median, of the PDF
  - variances and covariances of the parameters

- The choice of summary statistic depends on the situation:
  - In the case of non-Gaussian posteriors, one may want to characterize the posterior using quantiles, which can also illustrate potential asymmetries
  - In the case of a multi-modal posterior, one may want to report the location of the maxima

Exercise: Bayesian estimation – first try

- Suppose we want to measure the intensity λ of a light source by counting the number of photons, n, detected in a certain time interval. We assume that $n \sim$ Poisson(λ).

- Given $n = 10$ what is the estimate of λ?

$$\mathcal{L}(n \,|\, \lambda) = \frac{\lambda^n \exp(-\lambda)}{n!} \qquad \text{MLE:} \quad \tilde{\lambda} = n = 10$$

  (Exercise: can you show this?) →Blackboard

- The MAP in Bayesian estimation gives the same result if the prior is flat
  - for example, consider a prior that is flat between 0 and 100

- But if the prior state of knowledge is that we have no idea of the *order of magnitude* of λ then it can be argued that the cumulative probability of the prior should be flat in $\log\lambda$ (Why?) implying a PDF inversely proportional to λ meaning

$$\text{posterior PDF} \propto \mathcal{L}(n \,|\, \lambda)\, \lambda^{-1} = \exp(-\lambda)\, \lambda^{n-1}$$

Exercise: Bayesian estimation – first try

- So generally, one may write

$$\text{posterior } P(\lambda|n) \propto \mathcal{L}(n\,|\,\lambda)\,\lambda^k \propto \lambda^{n+k}\exp(-\lambda)$$

- Complete the following table giving the result of the Bayesian estimate of λ

prior (λ^k)	MAP	$\mathrm{E}(\lambda|n)$ (mean)
$\lambda^0 = \mathrm{const}$	n	?
λ^{-1}	?	?

- Note, that the priors we are using are improper, i.e. cannot be normalized

- The following auxiliary formula coming from the Γ-function helps

$$\int_0^\infty \lambda^n \exp(-\lambda)\,d\lambda = n!$$

Interlude: the quadratic approximation

- Posterior distributions over a model parameter $P(\theta|D)$ are often "peaky" around a single mode.

- Similar to the log likelihood, we want to consider a posterior $F \equiv \ln P(\theta|D)$ and Taylor-expand F in the vicinity of the mode $\tilde{\theta}$.

$$F(\theta) \approx F(\tilde{\theta}) + (\theta - \tilde{\theta}) \left.\frac{dF}{d\theta}\right|_{\tilde{\theta}} + \frac{1}{2}(\theta - \tilde{\theta})^2 \left.\frac{d^2F}{d\theta^2}\right|_{\tilde{\theta}}$$

  - Since $\left.\frac{dF}{d\theta}\right|_{\tilde{\theta}} = 0$ at the maximum (mode), the linear term drops out, simplifying the expression to:

$$F(\theta) \approx F(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^2 \left.\frac{d^2F}{d\theta^2}\right|_{\tilde{\theta}}$$

- The exponential of the above expansion for F, giving back P, is:

$$P(\theta|D) \approx A \exp\left(\frac{1}{2}(\theta - \tilde{\theta})^2 \left.\frac{d^2F}{d\theta^2}\right|_{\tilde{\theta}}\right) \quad \text{with } A \equiv \exp(F(\tilde{\theta}))$$

Interlude: the quadratic approximation

$$P(\theta|D) \approx A \exp\left(\frac{1}{2}(\theta - \tilde{\theta})^2 \left.\frac{d^2F}{d\theta^2}\right|_{\tilde{\theta}}\right) \quad \text{with } A \equiv \exp(F(\tilde{\theta}))$$

- This is a Gaussian distribution with mean $\tilde{\theta}$ and variance:

$$\sigma^2 = \left(-\left.\frac{d^2F}{d\theta^2}\right|_{\tilde{\theta}}\right)^{-1}$$

- The quadratic approximation simplifies the analysis by approximating the posterior distribution with a normal distribution centered at the mode $\tilde{\theta}$ with a variance given by the inverse of the second derivative of F

- The quadratic approximation can be very useful for large or complex models where exact analytical solutions are infeasible

Interlude: the quadratic approximation

- In multiple dimensions, this generalizes to (using the Hessian matrix):

$$\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\right] = \left(- \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \bigg|_{\tilde{\boldsymbol{\theta}}} \right)^{-1}$$

- Looks familiar if we associate F with $\ln \mathcal{L}$ (the log-likelihood)

- In Bayesian estimation, improper priors can lead to posteriors that cannot be normalized
  - In such cases, the variance cannot be computed directly

- By fitting a Gaussian to the peak of the posterior PDF, the quadratic approximation allows us to compute an *approximate* variance
  - The accuracy of this approximation depends on the problem
  - The posterior should not deviate too much from a Gaussian for this approximation to be valid

Interlude: transforming random variables

- Suppose we are given a PDF $P_X(x)$ of a random variable X. We subject x to a variable transformation $y = y(x)$
  We now ask: What is the PDF $P_Y(y)$ of the new random variable Y?

- The principle of *conservation of probability* states that the probability must be conserved under the transformation: $|P_X(x)\,dx| = |P_Y(y)\,dy|$ so that

$$P_Y(y) = P_X(x)\left|\frac{dx}{dy}\right| = P_X(x(y))\left|\frac{dx}{dy}\right| \qquad \text{(for a 1D transformation)}$$

- In j dimensions this generalizes to vectorial variables/functions:

$$P_Y(\mathbf{y}) = P_X(\mathbf{x}(\mathbf{y}))\left|\frac{\partial(x_1, ..., x_j)}{\partial(y_1, ..., y_j)}\right|$$

  Here, $\mathbf{x}$ and $\mathbf{y}$ are vectors, and $\left|\frac{\partial(y_1,...,y_j)}{\partial(x_1,...,x_j)}\right|$ is the determinant of the Jacobian matrix of the transformation

→ Example: normalized (standardized) Gaussian variables →Blackboard

Interlude: linear transformation of random vectors

- When you apply a linear transformation to a random vector $\mathbf{x}$, the resulting vector $\mathbf{y}$ can be expressed as: $\mathbf{y} = \mathbf{c} + \mathbf{A}\mathbf{x}$, where $\mathbf{c}$ is a fixed vector (constant shift) and $\mathbf{A}$ is a fixed matrix (which scales, rotates, or otherwise linearly transforms the vector $\mathbf{x}$

- The expectation vector $\mathrm{E}[\mathbf{y}]$ is given by: $\mathrm{E}[\mathbf{y}] = \mathbf{c} + \mathbf{A}\,\mathrm{E}[\mathbf{x}]$
  - "the expectation of the transformed vector $\mathbf{y}$ is the mean of $\mathbf{x}$ scaled by $\mathbf{A}$ and then shifted by $\mathbf{c}$"

- The covariance matrix of $\mathbf{y}$ is given by: $\mathrm{Cov}\left[\mathbf{y}, \mathbf{y}^{\mathrm{T}}\right] = \mathbf{A}\,\mathrm{Cov}\left[\mathbf{x}, \mathbf{x}^{\mathrm{T}}\right]\mathbf{A}^{\mathrm{T}}$
  - "the covariance of the transformed vector $\mathbf{y}$ is obtained by scaling the covariance of $\mathbf{x}$ by $\mathbf{A}$ on the left and by $\mathbf{A}^{\mathrm{T}}$ on the right"

- (Physicists may recognize this as similar to rotation and transformation to the principal axes of ellipsoids, which is a common technique in mechanics and quantum mechanics)

Example: Linear transformation of a random vector

Given:

- A random vector $\mathbf{x}$ with:
  - mean $\mathrm{E}[\mathbf{x}] = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and covariance matrix $\mathrm{Cov}\left[\mathbf{x}, \mathbf{x}^{\mathrm{T}}\right] = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$
- A linear transformation defined by:
  - $\mathbf{c} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$

Use R **to find:**

- The mean vector $\mathrm{E}[\mathbf{y}]$ and covariance matrix $\mathrm{Cov}\left[\mathbf{y}, \mathbf{y}^{\mathrm{T}}\right]$ of the transformed vector $\mathbf{y}$
- R Hints: use the `matrix` function and `%*%` to perform matrix multiplication. The transpose of matrix A is given by the command `t(A)`

Example: Linear transformation of a random vector

Solution:

1. **Mean Vector:**

$$\mathrm{E}[\mathbf{y}] = \mathbf{c} + \mathbf{A}\mathrm{E}[\mathbf{x}] = \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 4 \\ 9 \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$$

2. **Covariance Matrix:**

$$\mathrm{Cov}\left[\mathbf{y}, \mathbf{y}^{\mathrm{T}}\right] = \mathbf{A}\mathrm{Cov}\left[\mathbf{x}, \mathbf{x}^{\mathrm{T}}\right]\mathbf{A}^{\mathrm{T}} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} =$$

$$= \begin{pmatrix} 2 & 1 \\ 1.5 & 6 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 3 \\ 3 & 18 \end{pmatrix}$$

in R → `day7_example_linear_transform.ipynb`

Interlude: summing random variable

- Given two *independent* random variables X and Y, distributed as $P_X(x)$ and $P_Y(y)$, respectively, we can ask what is the distribution P_Z of the sum $Z \equiv X + Y$?

- Since X and Y are independent, we can write $P_Y(y|x) = P_Y(y)$

- By the law of total probability

$$P_Z(z) = \int_{-\infty}^{+\infty} P_Y(y = z - x|x)\, P_X(x)\, dx = \int_{-\infty}^{+\infty} P_Y(z - x)\, P_X(x)\, dx$$

  so that $P_Z = P_X * P_Y$, where "$*$" denotes the convolution of the two distributions

- Summary: When summing two independent random variables X and Y, the distribution of their sum Z can be found using the convolution of their individual distributions
  - The convolution can be interpreted as a "smoothing" operation that combines the two functions

Another example: Should Sam Guy play?

Let's say Mr Sam Guy is a basketball player who plays in the American NBA league. His performance over the last season was quite good by NBA standards in terms of his 3-point scores, namely having 79 scores in 209 shots taken But in the first quarter of the current season his performance has apparently dropped, with only 15 scores in 50 shots taken. If you were the coach of the team would you pick him to play in forthcoming matches, or would you start looking for a replacement? What does statistics tell you?

We want, of course, consider Mr Guy's scoring performance as a random experiment. Scoring or not-scoring is obviously a Bernoulli experiment, following binomial statistics. However, we are not so much interested in the actual number of points scored but rather (in the spirit of Thomas Bayes) in the PDF of Mr Guy's "scoring probability". Let's call it p. The question is this: what PDF of p can we expect of Mr. Guy for the rest of the current season, after he performed comparatively poorly during the first quarter?

Before you go into this, here's a story about *conjugate priors* and the `Beta` distribution . . .

→ blackboard

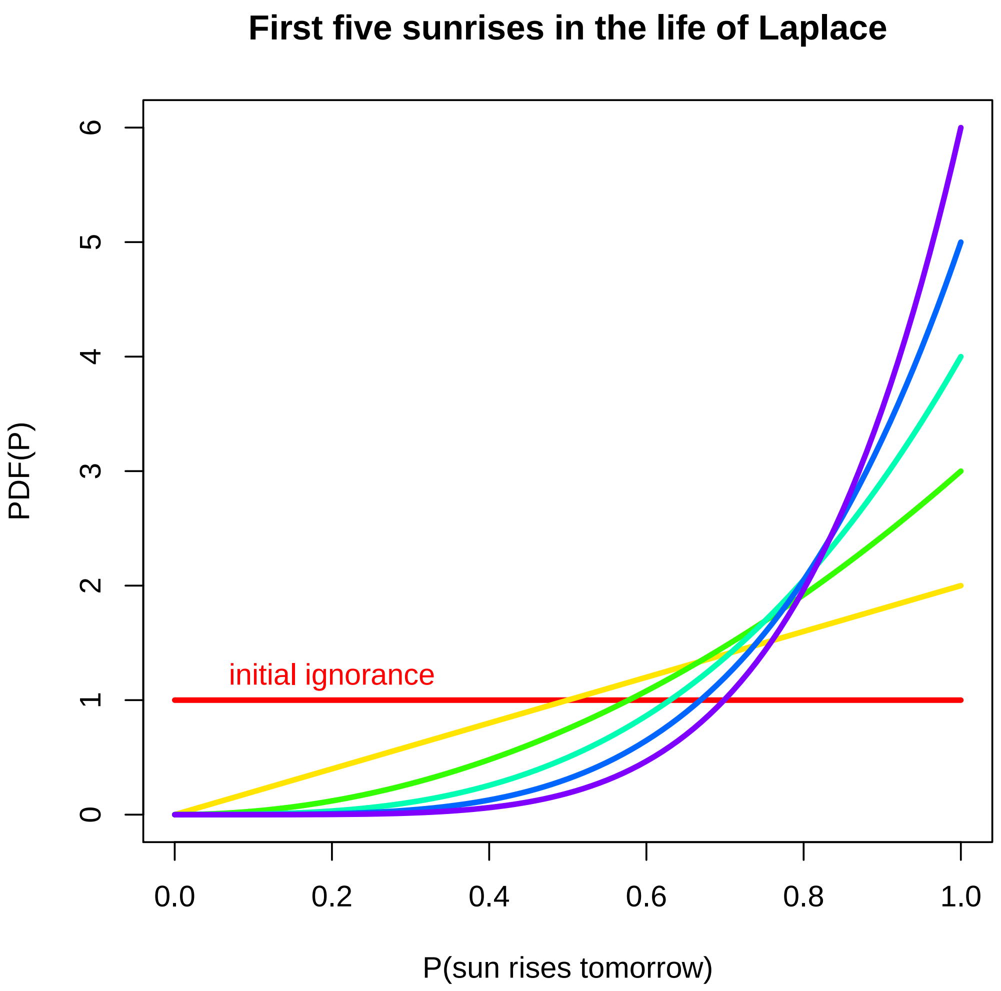
Exercise: Should Sam Guy play?

1. Consider Mr. Guy's scoring performance in the last and current season as separate random experiments, and plot the PDFs of p which describe his "scoring statistics" in the respective seasons. Assume an uninformative prior (using the β-distribution) in both cases.

2. Looking at Mr Guy's performance of the current season only: would you let him continue to play? Any hit-rate below 30 % is considered below NBA standards.

3. Now use the posterior of last season as prior for the current season. Based on the new combined posterior: would you let him play?

Laplace's sunrise problem



- Predicting the probability that the Sun is rising the next day from a purely statistical point of view
- The plot shows how Laplace's opinion evolves from day to day
- From complete ignorance (red line) to something more peaked towards $p = 1$ after having observed 5 sunrises (blue-violet line)
- Expectation $\mathrm{E}[p]$ perhaps more informative than MAP (always $p_{\mathrm{MAP}} = 1$) here

- Laplace's *rule of succession*: $\mathrm{E}[p] = \frac{1+k}{2+n}$ (k #successes, n #trials)