An overview of Bayesian inference Part I

Thomas Hamelryck

Bioinformatics center, University of Copenhagen

September, 2017

The information revolution and its problems

- We live in an era that produces enormous amounts of data.
  - Science: climate data, new generation sequencing of genomes, scientific literature
  - Society: government, health, digital archives
  - Business: e-commerce, social media, personalization
- We need tools that turn this data into knowledge.
  - Modelling, visualization, searching, prediction...
- This is the goal of machine learning, which ideally is based on sound probabilistic reasoning [3].

Probabilistic modelling

- The aim of probabilistic modelling is to develop a mathematical model that represents the data obtained from a certain system.
- This model makes use of the mathematics of probability theory to represent uncertainties and noise.
- The Bayesian calculus provides the necessary machinery to infer the parameters of such a model from the data and to apply the model to problems of inference.

Bayesian view of probability

- The Bayesian view interprets probability as a *degree of belief* or a *measure of certainty*.
- An alternative is the *frequentist view*, which interprets probability as a frequency.
  - An event's probability is the frequency of observing that event in a large number of trials.
- The Bayesian view has many advantages
  - Has a firm axiomatic base.
  - Makes intuitive sense.
  - Can be applied to a wide range of problems.
  - Can be implemented efficiently on computers.
  - Is general, ie. as opposed to a collection of *ad hoc* methods.

Laplace and Stigler's law

- Bayesian statistics derives its name from **Thomas Bayes** (170?-1761), an English Presbyterian minister who proved a special case of what is now called "Bayes' theorem".
- However, it was the French mathematician **Pierre-Simon Laplace** (1749-1827) who actually formulated the general case, and used it to solve problems in many areas of science.
- A classic case of what is called Stigler's Law.
  - "No scientific discovery is named after its original discoverer" (Note: Stigler was not the first to come up with this law)
- No doubt, those skeptical about Bayesian statistics found it easier to argue against an obscure Presbyterian minister than against the genius of Laplace.

Bayes' picture

- Bayes' alleged picture on the left is an icon in statistics.
- However, this is not Thomas Bayes, as both wig and clothes are from the wrong time period [1]. It's clearly Charlie Sheen. The picture on the right indeed shows Laplace.

The rise of frequentist statistics

- For about 100 years, the Bayesian interpretation reigned supreme.
- In the beginning of the 20th century, the so-called **frequentist interpretation** emerged.
  - Ronald Fisher, Jerzy Neyman, Egon Pearson
- The philosophical difference is illustrated by the "sunrise problem"
  - What is the probability that the sun will rise tomorrow?
  - The Bayesian approach formulates a model, estimates its parameters using the Bayesian calculus and uses the model to calculate the probability.
  - From a frequentist point of view, the question is meaningless, as there is no way to calculate this probability as a frequency in a large number of trials.

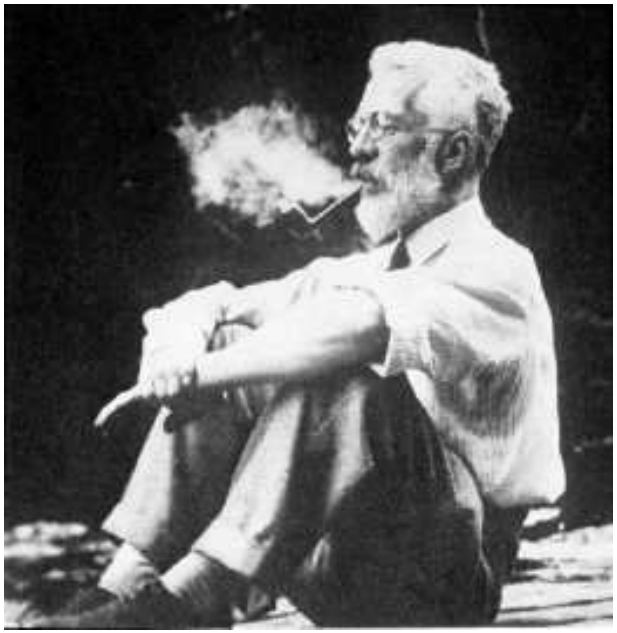
The theory that would not die

- During the second half of the 20th century, Bayesian statistics made a strong comeback [8].
  - The frequentist approach is plagued by inconsistencies and limitations.
  - Bayesian models are often analytically intractable and thus require methods based on simulation. Cheap and fast computers, and general-purpose software such as BUGS, resolved this issue.
    - Probabilistic programming (STAN, pyMC3) is the next leap forward.
  - Hierarchical graphical models, such as hidden Markov models and Bayesian networks, provided a unified framework for Bayesian computation.
  - Bayesian approaches are very common in machine learning.

Bayes at war

- In WW2, German U-boats communicated using a code generated by the Enigma machine.
- Alan Turing cracked this code using Bayesian methods. It wasn't until 1973 that the story of Turing and Bayes began to emerge [8].

Bayes and smoking



- In 1951, Jerome Cornfield used Bayesian methods to uncover the link between smoking and lung cancer.
- Frequentists Fisher and Neyman disagreed for many years.
- In 1959, Cornfield published a paper that systematically addressed Fisher's arguments. Fisher and his methods lost a lot of credibility [8].

Sum and product rule

- Probability theory builds on two fundamental rules.
- Sum rule
  - Discrete case

$$p(a) = \sum_b p(a, b)$$

  - Continuous case

$$p(a) = \int p(a, b) \mathrm{d}b$$

- Product rule

$$p(a, b) = p(a \mid b)p(b) = p(b \mid a)p(a)$$

Bayes' theorem

- **Bayes' theorem** follows directly from the product rule

$$
\begin{aligned}
p(a \mid b)p(b) &= p(b \mid a)p(a) \\
&\Rightarrow \\
p(a \mid b) &= \frac{p(b \mid a)p(a)}{p(b)} \\
\text{and } p(b \mid a) &= \frac{p(a \mid b)p(b)}{p(a)}
\end{aligned}
$$

- Bayes theorem is perfectly valid in both Bayesian and frequentist statistics.
- The Bayesian calculus is more than just an application of Bayes' theorem.
  - See part II, foundations.

Conditional independence

- a is **conditionally independent** of b given c, if

$$\begin{aligned} p(a \mid b, c) &= p(a \mid c) \\ p(a, b \mid c) &= p(a \mid c)p(b \mid c) \end{aligned}$$

- Conditional independence plays an important role in formulating tractable hierarchical probabilistic models.
- Two **independent random variables** are conditionally independent given the empty set.

$$\begin{aligned} p(a, b) &= p(a)p(b) \\ p(a, b \mid \emptyset) &= p(a \mid \emptyset)p(b \mid \emptyset) \end{aligned}$$

Change of variables - Jacobian

- Given a density $p_X(x)$, suppose there is a **change of variables** $y = f(x)$. What is $p_Y(y)$? The solution is

$$p_Y(y) = p_X(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| = \frac{p_X(x)}{|f'(x)|}.$$

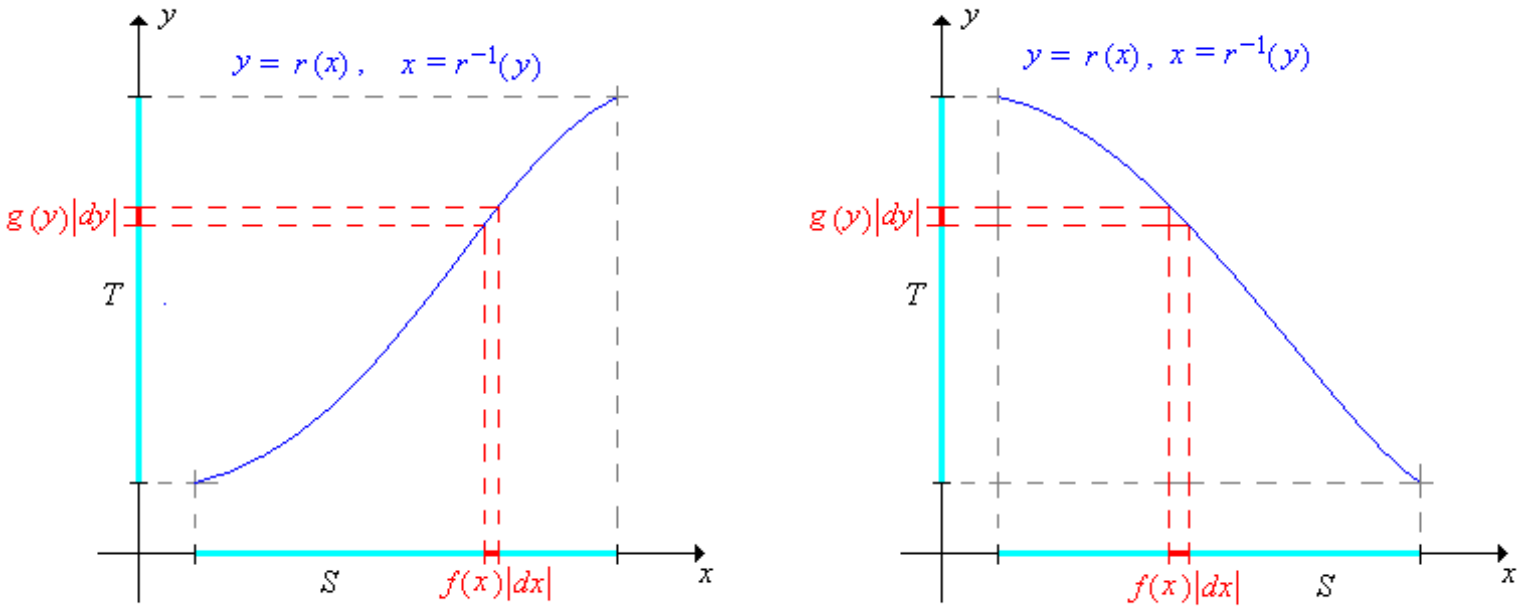
- The absolute value ensures that the density is positive.

Given $p_X(x)$ with $x \geq 0$ and $y = f(x) = x^2$, what is $p_Y(y)$?

- We know $|f'(x)| = 2x$ and $x = \sqrt{y}$, and thus

$$\frac{p_X(x)}{|f'(x)|} = \frac{p_X(x)}{2x} \Rightarrow p_Y(y) = \frac{p_X(\sqrt{y})}{2\sqrt{y}}.$$

The Jacobian explained



- The factor $\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$ is called the Jacobian, after the German mathematician Karl Gustav Jacobi (1804–1851)[1].
- The probability in a differential area must be invariant under change of variables, thus $p(y)\,|\mathrm{d}y| = p(x)\,|\mathrm{d}x|$.

[1] Picture: http://www.randomservices.org/random/dist/Transformations.html

Conditional probability tables

- For discrete random variables with a finite number of states.
- CPTs specify the probability of one variable conditioned on one or more variables.

Example: $P(\text{car ownership} \mid \text{size of income})$

	No Car	Second hand car	New car
Low income	0.2	0.5	0.3
High income	0.1	0.2	0.7

The Dirichlet distribution

- The Dirichlet distribution is a probability density over k-dimensional probability vectors $\boldsymbol{\mu}$.
  - Vector with k positive components, that sum to 1
  - Devised by the German mathematician Johann Peter Gustav Lejeune Dirichlet (1805-1859).
  - The support of the Dirichlet distribution is the k-dimensional simplex, which is a generalization of a triangle.
  - This is an example of a probability distribution on a "special" space. We'll see more of that later.
  - The Dirichlet distribution is often used as a prior distribution for a probability vector.

The Dirichlet distribution

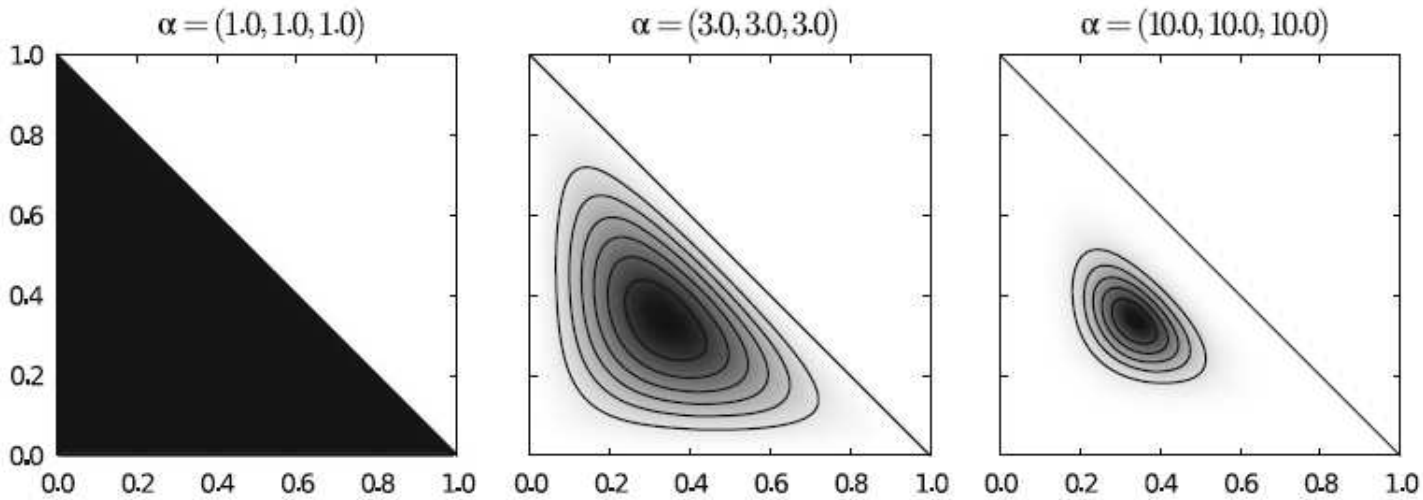
- The PDF is

$$p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{1}{C} \prod_{i=1}^{k} \mu_i^{\alpha_i - 1}$$

  where $\boldsymbol{\alpha}$ is a k-dimensional vector of positive real numbers.
  - C is a normalization factor.
- The two-dimensional version of the Dirichlet distribution is called the Beta distribution.

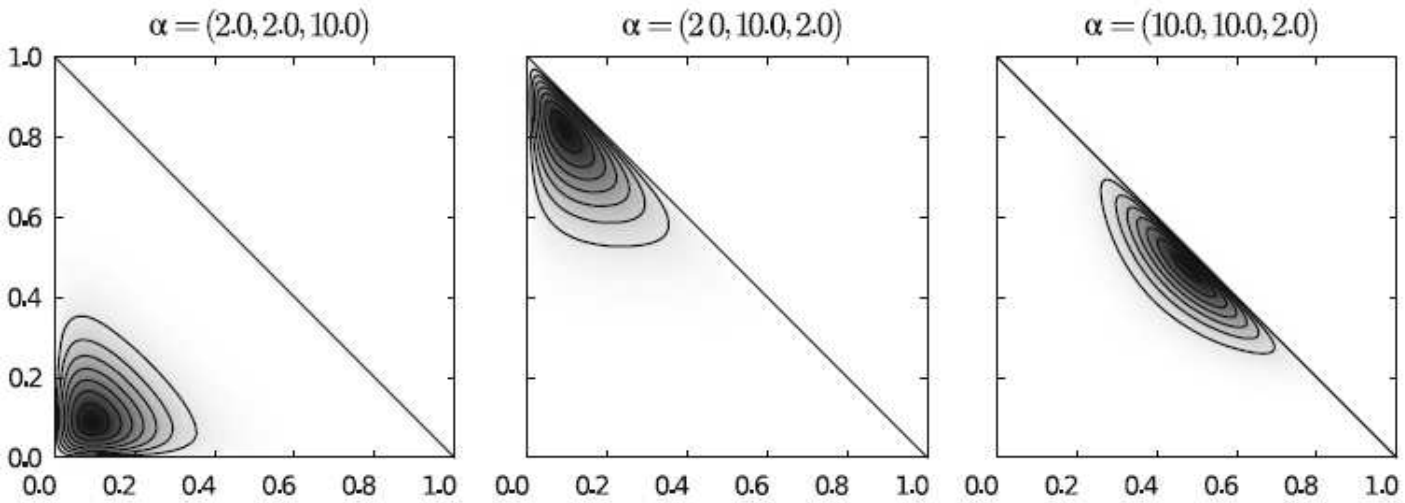
The Dirichlet distribution

- Some 3-dimensional Dirichlet distributions.



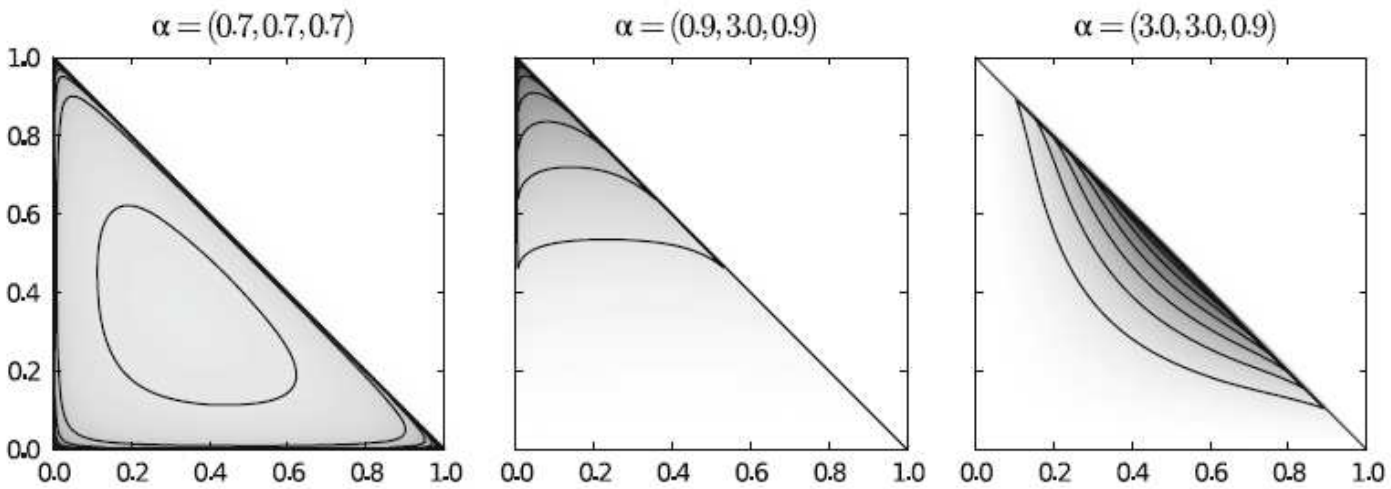
The Dirichlet distribution

- Some 3-dimensional Dirichlet distributions.



The Dirichlet distribution

- Some 3-dimensional Dirichlet distributions.



Directional statistics I

- For data in Euclidean space, one often uses the multivariate Gaussian distribution.
- But there are other spaces (manifolds) than Euclidean space.
  - Torus, sphere, cylinder, projective space,...

Question

What would you use for data on the positive real axis, $[0, +\infty]$? Note that the Gaussian has support $[-\infty, +\infty]$.

- You can transform your way out of it.

Directional statistics I

- For data in Euclidean space, one often uses the multivariate Gaussian distribution.
- But there are other spaces (manifolds) than Euclidean space.
  - Torus, sphere, cylinder, projective space,...

Question

What would you use for data on the positive real axis, $[0, +\infty]$? Note that the Gaussian has support $[-\infty, +\infty]$.

- You can transform your way out of it.
- Take the log of the data, so the data is in $[-\infty, +\infty]$.

Directional statistics I

- For data in Euclidean space, one often uses the multivariate Gaussian distribution.
- But there are other spaces (manifolds) than Euclidean space.
  - Torus, sphere, cylinder, projective space,...

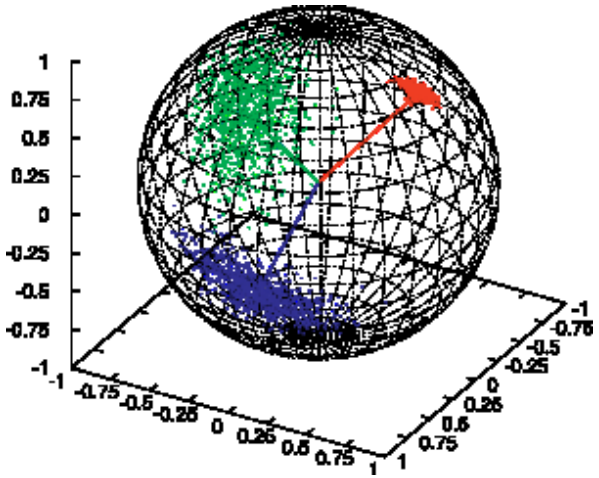
Question

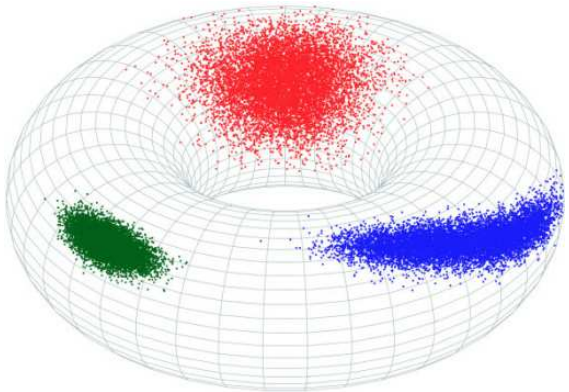
What would you use for data on the positive real axis, $[0, +\infty]$? Note that the Gaussian has support $[-\infty, +\infty]$.

- You can transform your way out of it.
- Take the log of the data, so the data is in $[-\infty, +\infty]$.
- Use the Gaussian distribution.

Directional statistics II

- Sometimes no transformation will work for the general case – you need distributions on the right manifold.
  - This is the realm of directional statistics.
- For example, the Kent and bivariate von Mises distributions on the sphere and the torus. They are used for modelling 3D protein structure [4, 6].





Full Bayesian approach I

- The goal of Bayesian inference is to obtain the probability distribution over the parameters $\mathbf{h}$ of a probabilistic model given the data $\mathbf{d}$.
- This is called the posterior distribution.
- Note that we obtain a probability distribution over all possible values of the parameters, rather than an "optimal" point estimate.
- In order to calculate the posterior, we need:
  - the likelihood $p(\mathbf{d} \mid \mathbf{h})$, which brings in the data.
  - the prior $p(\mathbf{h})$, which specifies the situation before the data was observed.
  - the evidence $p(\mathbf{d})$, which is the marginal probability of the data.

Full Bayesian approach II

- The Bayesian calculus is simple

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

$$p(\mathbf{h} \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid \mathbf{h})p(\mathbf{h})}{p(\mathbf{d})}$$

with $p(\mathbf{d}) = \int p(\mathbf{d} \mid \mathbf{h})p(\mathbf{h})\mathrm{d}\mathbf{h}$.

- The evidence is constant, and can often be ignored.

$$p(\mathbf{h} \mid \mathbf{d}) \propto p(\mathbf{d} \mid \mathbf{h})p(\mathbf{h})$$

Example: The binomial distribution I

- Let's consider Bayesian inference of the parameter θ of the binomial distribution.
  - Model for trials with binary outcome - "success" or "failure".
  - Probability of observing k successes in n trials is given by

$$p(k \mid \theta, n) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}$$

  - **Note**: n is given as part of the experimental design.
- Given k and n, how do we infer θ according to the rules of the Bayesian calculus?

Example: The binomial distribution II

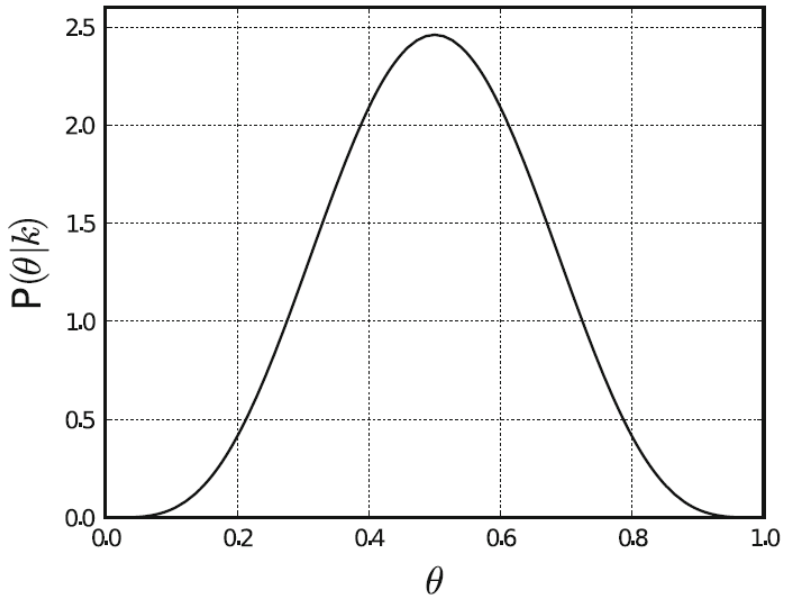
- Following, posterior $\propto$ likelihood $\times$ prior, we get

$$p(\theta \mid k, n) \propto p(k \mid \theta, n)\pi(\theta)$$

- Obviously, the first factor is just the binomial distribution.
- If we chose a uniform distribution for the prior $\pi(\theta)$, we obtain

$$p(\theta \mid k, n) \propto \theta^k (1-\theta)^{n-k}$$

Example: The binomial distribution III



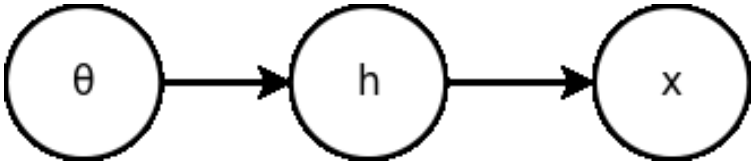
- The posterior $\theta^k(1-\theta)^{n-k}$ corresponds to the Beta distribution, which is a 2-dimensional Dirichlet distribution.
- As expected, the posterior peaks at 0.5 for $k = n - k = 5$.
- One might think that the uniform prior represents "complete ignorance", but this is not the case. The ignorance prior is the Jeffrey's prior.

Sequential application of the Bayesian calculus

- The posterior from one data set $\mathbf{d}$ can serve as the prior when another data set $\mathbf{d}'$ becomes available.
  - We assume the data sets are conditionally independent given $\mathbf{h}$
- Below we first apply the product rule and then invoke the conditionally independence assumption.

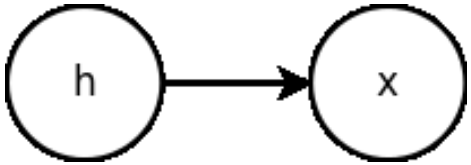
$$\begin{aligned} p(\mathbf{h} \mid \mathbf{d}, \mathbf{d}') &\propto p(\mathbf{d}, \mathbf{d}' \mid \mathbf{h}) \pi(\mathbf{h}) \\ &= p(\mathbf{d}' \mid \mathbf{d}, \mathbf{h}) p(\mathbf{d} \mid \mathbf{h}) \pi(\mathbf{h}) \\ &= p(\mathbf{d}' \mid \mathbf{h}) p(\mathbf{d} \mid \mathbf{h}) \pi(\mathbf{h}) \\ &= \text{likelihood from } \mathbf{d}' \times \text{posterior from } \mathbf{d} \end{aligned}$$

Hierarchical models and nuisance variables



- Bayesian models often have a hierarchical structure. Observations depend on parameters which depend on other parameters – called *hyperparameters*. The chain ends at the priors with given, fixed hyperparameters.
- These models often contain *nuisance variables* – these are necessary to construct a valid model but are not of interest themselves.
- Nuisance variables are often unobserved or *latent.*
- A classic example of a hierarchical model is the *mixture model.*

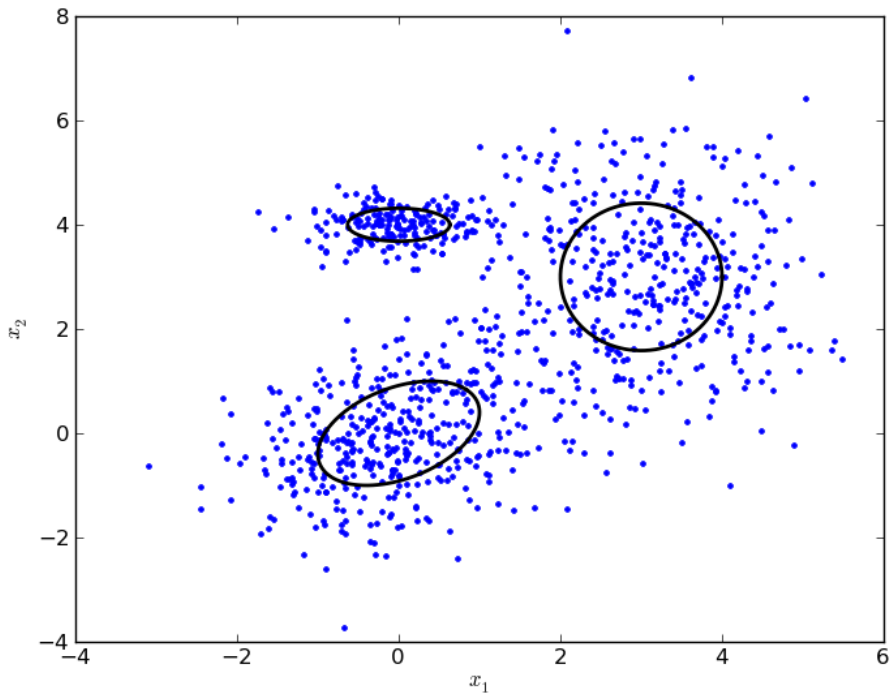
The Gaussian mixture model



- h adopts a finite number H of discrete values, each with an associated probability. H is the number of *mixture components*.
- $p(x \mid h)$ is given by the Gaussian $p(x \mid \mu_h, \sigma_h)$. The mixture component h specifies the mean and standard deviation.
- Mixture models allow modelling multimodal distributions using standard unimodal distributions.

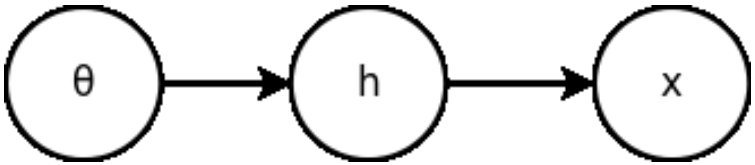
$$p(x) = \sum_{h=1}^{H} \mathcal{N}(x \mid \mu_h, \sigma_h) p(h)$$

Gaussian mixture model: example



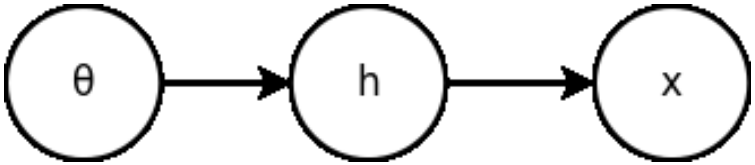
- A 2D Gaussian mixture model with three components. The ellipses are equiprobability contours.

Bayesian inference of mixture models



- $p(h)$ is specified by a probability vector $\boldsymbol{\theta}$. Following the Bayesian calculus, we want a density over $\boldsymbol{\theta}$. Hence, an additional node is added to the hierarchical model.
- For inference of $\boldsymbol{\theta}$, we need to specify a prior $p(\theta)$ with fixed parameters – a Dirichlet distribution is the classic choice.
- Already for this simple model, the posterior cannot be written down analytically. Estimation and inference can however be done using approaches based on sampling.

Graphical models and Bayesian networks



- The mixture model is an example of a graphical model.
- Because the arrows are directed and there are no cycles, it's a Bayesian network.

$$p(\mathbf{z}) = \prod_i p(z_i \mid \mathrm{Parents}(z_i))$$

- Graphical models are carriers of *conditional independencies*. If two nodes are *not* connected by an arrow, they are guaranteed to be conditionally independent give a third set of nodes. Hence,

$$p(x \mid \theta, h) = p(x \mid h)$$

Monte Carlo and Bayes

- The posterior is often unavailable in closed analytic form. The mixture model is a good example.
- The posterior can be approximated using methods based on sampling – Monte Carlo methods [5]. Fast computers made this approach possible!
- The core idea is simple – any expectation $\mathbb{E}[f(x)]$ can be approximated by sampling,

$$\mathbb{E}[f(x)] = \int f(x)p(x)\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

with $x_s \sim p(x)$ and S is the number of samples.

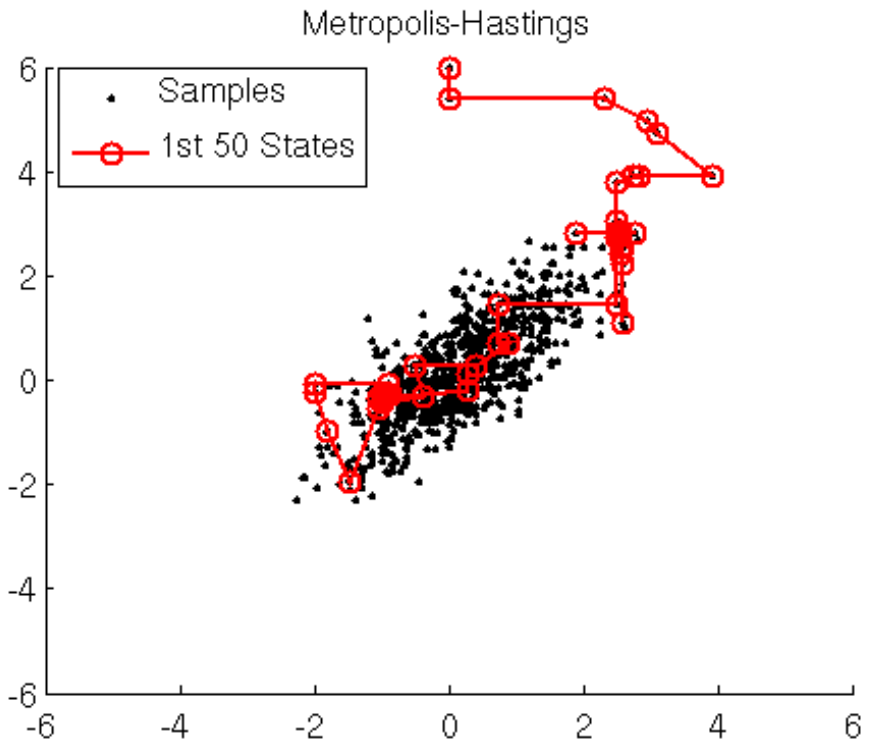
Markov chain Monte Carlo methods

- The idea is to construct a biased random walk that explores the target distribution $p(x)$.
  - This idea was launched by Gelfand and Smith in 1990.
- MCMC methods generate approximate - *but correlated* - samples from $p(x)$.
- The emergence of fast computers made MCMC methods tractable. This is a main reason behind the rapid spread of Bayesian methods in the late 20th century.
- Many methods exist
  - Metropolis-Hastings and Metropolis sampling
  - Gibbs sampling

Markov chains

- We generate a sequence of random variables $\{x_0, x_1, \ldots\}$ such that at each time $t \geq 0$, the next variable x_{t+1} is sampled from a distribution $p(x_{t+1} \mid x_t)$.
  - Markov property: the history of the chain before x_t does not matter.
  - $p(x_{t+1} \mid x_t)$ is the *transition kernel*.
- Under general conditions, the chain will "forget" the initial state x_0 with time.
  - $p(x_t \mid x_0)$ will converge to the *stationary* or *invariant distribution* for $t \to \infty$.
- The idea is to construct a Markov chain whose stationary distribution is the posterior of interest.

MCMC sampling of a bivariate Gaussian



- Metropolis-Hastings sampling[2]

[2] Picture from https://theclevermachine.wordpress.com/tag/hybrid-monte-carlo/

Metropolis-Hastings sampling

Algorithm: Markov chain with stationary distribution p(x)

- Start with a random point x
- Repeat until enough samples have been generated:
  - Generate a potential next sample x' from a proposal distribution $q(x' \mid x)$
  - Accept x' with probability $\min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right)$.
  - Otherwise, set x' to x (re-use x as next sample)

- The choice of the proposal distribution is critical.
- First samples (from the burn-in period) are discarded.
- In *Metropolis sampling*, $q(x' \mid x) = q(x \mid x')$.
  - For example, $q(x' \mid x)$ could be a Gaussian $\mathcal{N}(x' \mid x, \sigma)$.

Gibbs sampling

- For multivariate probability distributions.
- A variant of MH-MCMC. All samples are accepted.
- Makes use of sampling from the conditional distributions.

Algorithm

- Chose a random start $\mathbf{x}$
  - $\mathbf{x}$ is an N-dimensional vector.
- Repeat until enough samples:
  - For every i, replace x_i with $x_i' \sim p(x_i' \mid \overbrace{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N}^{\text{latest samples without } x_i})$, where the conditioning is on the latest samples.

Point estimates

- The Bayesian posterior contains all information of interest on the parameters of interest. However, sometimes we might settle for a *point estimate*.
  - It might be a lot faster, while still offering a fairly good estimate of a sharply peaked posterior.
  - We might be dealing with a *decision problem*, where an optimal decision needs to be made.
- Point estimates such as maximum likelihood estimates are typically seen as "frequentist methods", but they are perfectly acceptable as approximation of full Bayesian inference.
  - These approximations WILL be very poor in some cases,however.

Decision problems and loss functions

- Decision problems require the definition of a loss function, $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$, which measures the cost of using point estimate $\widehat{\boldsymbol{\theta}}$ instead of the "true" value $\boldsymbol{\theta}$.
- The *expected loss* is

$$\bar{L}(\widehat{\boldsymbol{\theta}}) = \int L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{d}) \mathrm{d}\boldsymbol{\theta}$$

- A point estimate $\widehat{\boldsymbol{\theta}}$ that minimizes $\bar{L}(\widehat{\boldsymbol{\theta}})$ is a *Bayes estimator* for a given loss function.
- Different loss functions give rise to different point estimates. MAP and ML are the most common point estimates.

The MAP estimate

- The **maximum a posteriori** estimate uses the maximum (mode) of the posterior.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{d} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

- The MAP estimate optimizes the zero-one loss function.

$$\begin{aligned} L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= 0, \text{ if } \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\| \leq \varepsilon \\ L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= 1, \text{ if } \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\| > \varepsilon \end{aligned}$$

with $\epsilon \to 0$.

Different loss functions and their point estimates

- Mean, median and mode of the posterior all arise as Bayes estimators for specific loss functions.
- Mode for the zero-one error (L0)

$$\begin{aligned} L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= 0, \text{ if } \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\| \leq \varepsilon \\ L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= 1, \text{ if } \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\| > \varepsilon \end{aligned}$$

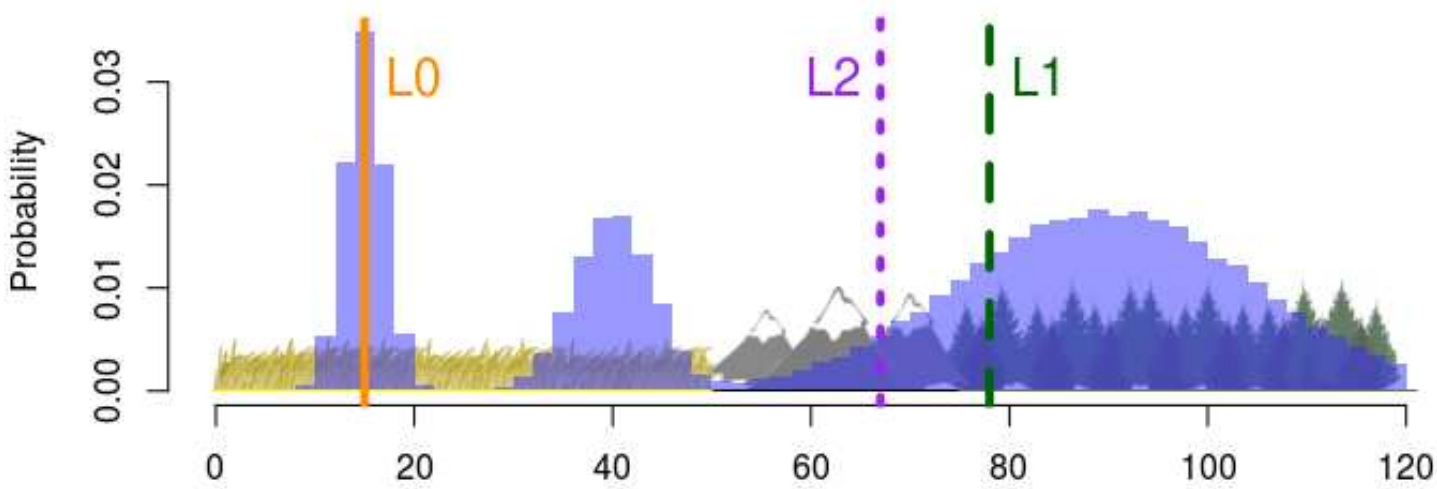
- Median for the absolute error (L1)

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|$$

- Mean for the quadratic error (L2)

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|^2$$

Search problem example I

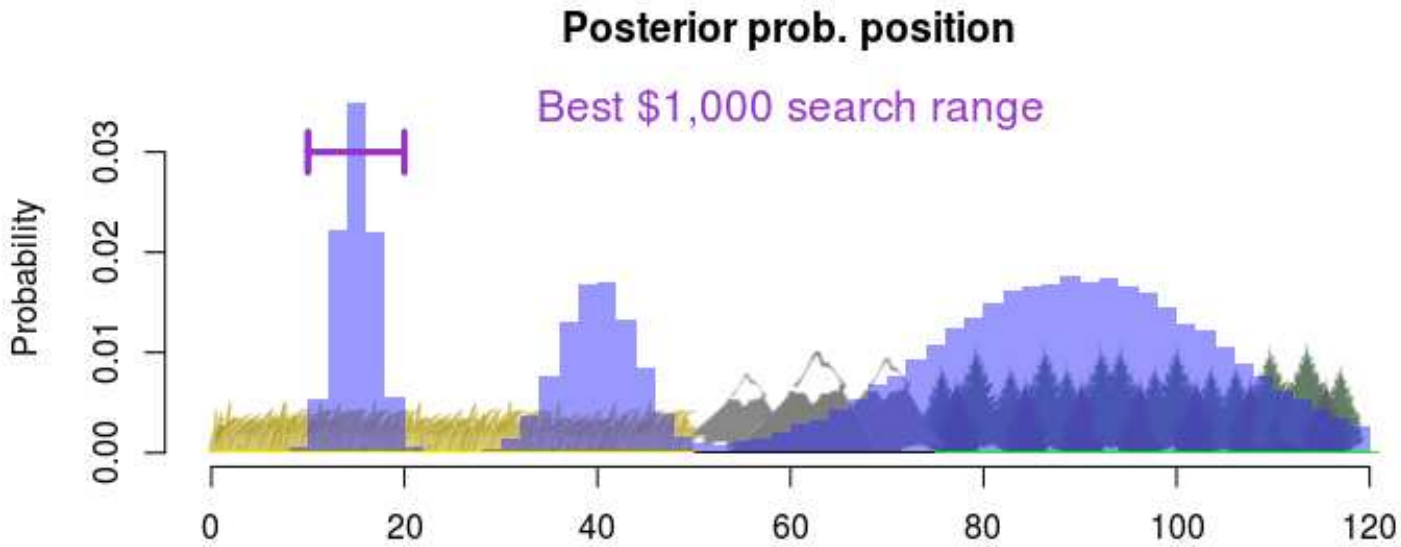


- The posterior (blue) specifies where a missing person might be located. L0, L1 and L2 indicate the best guess (mode, median and mean) for different loss functions.
- Picture from http://www.sumsar.net/blog/2015/01/probable-points-and-credible-intervals-part-two/

Bayesian credible interval

- Instead of a point estimate, we can summarize the posterior with an interval.
- Bayesian 95% **credible** interval
  - "I believe the value of the parameter of interest lies in that interval with probability 0.95."
    - Example: If θ has a posterior $\theta \sim \mathcal{N}(0, \sigma)$ then $[-1.96\sigma, 1.96\sigma]$ is a 95% Bayesian credible interval.
  - Intuitive interpretation.
- Frequentist 95% **confidence** interval
  - "For 100 experiments and the procedure chosen, at least 95 of the resulting confidence intervals will be expected to include the true value of the parameter."
  - Often interpreted wrongly as a Bayesian credible interval!

Search problem example II



- Now, say the search should cost maximum 1000$. We know the cost to search a mile of each type of terrain. Hence, we can pick the interval that costs 1000$ and that maximizes the probability of success.

The ML estimate

- If one assumes the prior is uniform and a zero-one error, one obtains the maximum likelihood estimate.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\theta} p(\mathbf{d} \mid \boldsymbol{\theta})$$

- The ML estimate is very common, and often used in frequentist methods.
- Often a good approximation but can go badly wrong.
  - If the data are sparse.
  - If a uniform prior actually induces strong and unsuited prior beliefs. A uniform prior is **not** necessarily a prior that indicates ignorance.

Asymptotic properties of point estimates

- Under some general assumptions, and for large data sets, one can approximate the posterior as a Gaussian with mean equal to the the ML estimate $\widehat{\theta}_{\mathrm{ML}}$.
- The variance is

$$\sigma^2 = \frac{1}{M\mathcal{I}(\widehat{\theta}_{\mathrm{ML}})}$$

  where $M \to \infty$ is the number of observations.
- For a density $p(x \mid \theta)$, the *Fisher information* is given by

$$\mathcal{I}(\theta) = -\mathbb{E}_{p(x|\theta)}\left[\frac{\mathrm{d}^2 \log p(x \mid \theta)}{\mathrm{d}\theta^2}\right]$$

  with $0 \leq \mathcal{I}(\theta) < \infty$.
- This generalizes to the multivariate case.

Interpretation of the Fisher information

- Consider the derivative of the log-likelihood function

$$\mathcal{L}(x \mid \theta) = \frac{\mathrm{d} \log p(x \mid \theta)}{\mathrm{d}\theta}$$

  - If $\mathcal{L}(x \mid \theta) \approx 0$, the data x does not provide much information on θ.
  - If $|\mathcal{L}(x \mid \theta)|$ or $\mathcal{L}(x \mid \theta)^2$ is large, the data x provides much information on θ.
  - Thus, $\mathcal{L}(d \mid \theta)^2$ can serve to measure the amount of information associated with x.

- Now, the Fisher information can be written as the expected value of the "information" $\mathcal{L}(d \mid \theta)^2$.

$$\mathcal{I}(\theta) = \mathbb{E}_{p(x|\theta)} \left[\left(\frac{\mathrm{d} \log p(x \mid \theta)}{\mathrm{d}\theta} \right)^2 \right].$$

Fisher information of the Gaussian I

- We use the following definition of the Fisher information

$$\mathcal{I}(\theta) = \mathbb{E}_{p(x|\theta)} \left[\left(\frac{\mathrm{d} \log p(x \mid \theta)}{\mathrm{d}\theta} \right)^2 \right].$$

- Let's consider the case of the Gaussian density with unknown mean θ and known variance σ^2. Thus,

$$\log p = -\frac{1}{2} \log 2\pi - \log \sigma - \frac{1}{2} \left(\frac{x-\theta}{\sigma^2} \right)^2 \Rightarrow \frac{\mathrm{d} \log p}{\mathrm{d}\theta} = \frac{x-\theta}{\sigma^2}.$$

- Hence

$$\mathcal{I}(\theta) = \int \underbrace{\left(\frac{x-\theta}{\sigma^2} \right)^2}_{(\mathrm{d} \log p / \mathrm{d}\theta)^2} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left(\frac{x-\theta}{\sigma} \right)^2 \right)}_{\text{Gaussian density}} \mathrm{d}x = \frac{1}{\sigma^2}$$

Fisher information of the Gaussian II

- Suppose we obtain 10^5 data points from a Gaussian with unknown mean θ and given variance σ^2.
- The ML estimate of the mean is the data average, $\widehat{\theta}_{\mathrm{ML}}$.
- Now, let's approximate the Bayesian posterior using the Fisher information. We know that for this case

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2}.$$

- Thus,

$$\text{posterior of } \theta \approx \mathcal{N}(\widehat{\theta}_{\mathrm{ML}}, \frac{\sigma^2}{10^5}),$$

  as expected.
- Will this be a good approximation?

Empirical Bayes estimation

- In Bayesian inference, the parameters of the prior distribution (α) are supposed to be given and known.

$$p(h \mid d) \propto p(d \mid h)\pi(h \mid \alpha)$$

- But what if that is not the case? One can use a point estimate for the parameters of the prior.

$$\widehat{\alpha}_{\mathrm{EB}} = \arg\max_{\alpha} p(d \mid \alpha) = \int p(d \mid h)\pi(h \mid \alpha)\mathrm{d}h$$

  - This *empirical Bayes estimate* corresponds to the maximum likelihood estimate of a hierarchical model.

- The result is an Empirical Bayes posterior.

$$p(h \mid d) \propto p(d \mid h)\pi(h \mid \widehat{\alpha}_{\mathrm{EB}})$$

Shrinkage estimators

- Suppose $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ are samples from an N-dimensional Gaussian with $N > 2$ and given covariance matrix $\sigma^2 \mathbf{I}_N$.
  - The ML estimate of the mean $\widehat{\boldsymbol{\mu}}_{\mathrm{ML}}$ is $\overline{\mathbf{y}}$.
- In 1955, Charles Stein showed that this ML estimate is suboptimal in terms of the expected squared error loss function for dim > 2.

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|^2$$

  - The superior James-Stein **shrinkage estimate** of the mean is

$$\widehat{\boldsymbol{\mu}}_{\mathrm{JS}} = \left(1 - \frac{\sigma^2 (N-2)}{\left\| \overline{\mathbf{y}} \right\|^2} \right) \overline{\mathbf{y}}$$

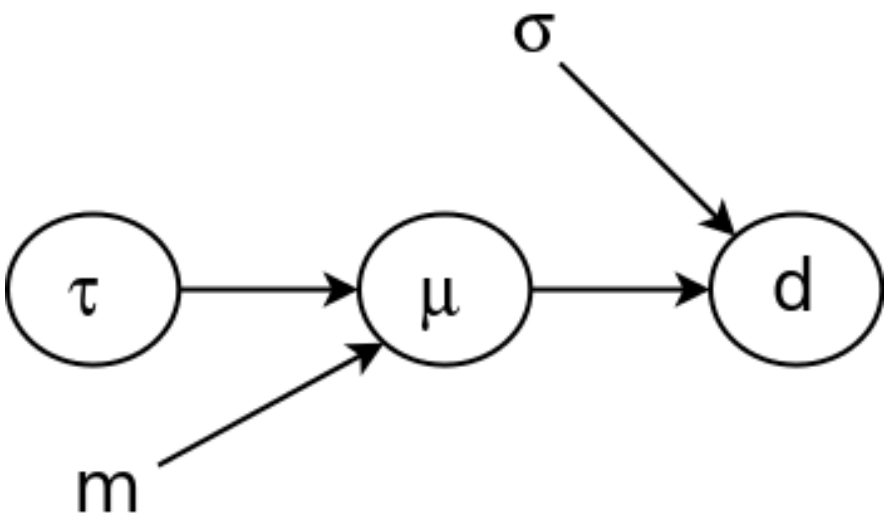
Bayesian view of shrinkage estimators

- A shrinkage estimator can be seen as a Bayes estimator.
  - Assume a prior over $\boldsymbol{\mu}$ that is Gaussian with mean $\mathbf{0}$ and covariance matrix $\tau^2 \mathbf{I}_N$, with τ unknown.
  - Estimate τ using the Empirical Bayes approach.

$$\begin{aligned} \widehat{\tau}_{\mathrm{EB}} &= \arg\max_{\tau} p(\mathbf{d} \mid \tau) \\ &= \arg\max_{\tau} p \int \mathcal{N}(\mathbf{d} \mid \boldsymbol{\mu}, \underbrace{\sigma^2 \mathbf{I}_N}_{\text{Known}}) \underbrace{\mathcal{N}(\boldsymbol{\mu} \mid \mathbf{0}, \tau^2 \mathbf{I}_N)}_{\text{Prior}} \mathrm{d}\boldsymbol{\mu} \end{aligned}$$

  - The JS estimate is the Bayes estimator of $\boldsymbol{\mu}$ under a squared loss function and using the estimate $\widehat{\tau}_{\mathrm{EB}}$.
- The prior "**shrinks**" the estimate of $\boldsymbol{\mu}$ towards $\mathbf{0}$, compared to the ML estimate. In this case, the uniform ML prior is overly informative.

Graphical model of JS-estimator



- The posterior (with given parameters underlined) is

$$p(\mathbf{d} \mid \boldsymbol{\mu}, \underline{\sigma}, \underline{\mathbf{m}}, \tau) \propto \underbrace{\mathcal{N}(\mathbf{d} \mid \boldsymbol{\mu}, \sigma^2 \mathbf{I}_N)}_{\text{likelihood}} \underbrace{\mathcal{N}(\boldsymbol{\mu} \mid \mathbf{m} = \mathbf{0}, \tau^2 \mathbf{I}_N)}_{\text{prior}}.$$

- The value of τ is determined using Empirical Bayes, $\widehat{\tau}_{\mathrm{EB}} = \arg\max_{\tau} p(\mathbf{d} \mid \tau)$.

Expectation maximization

- The goal of EM is ML estimation of $\boldsymbol{\theta}$ in the presence of latent nuisance parameters $\mathbf{h}$.
- The hidden parameters are integrated or summed away.

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{h}} p(\mathbf{d}, \mathbf{h} \mid \boldsymbol{\theta})$$

- This can be done using an iterative algorithm.
  - E-step: the hidden variables are estimated
  - M-step: update the estimate of $\boldsymbol{\theta}$
  - Guaranteed to converge to a local maximum.
- EM is often used for ML estimation of mixture models and hidden Markov models (the Baum-Welch algorithm).

EM algorithm

EM algorithm

- Start with an estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$
- E-step: infer the distribution of $\mathbf{h}$ for the current estimate of $\boldsymbol{\theta}$.

$$p(\mathbf{h} \mid \mathbf{d}, \widehat{\boldsymbol{\theta}})$$

- M-step: replace $\widehat{\boldsymbol{\theta}}$ with $\widehat{\boldsymbol{\theta}}_{\text{new}}$ which maximizes the *expectation* of the "completed" log-likelihood.

$$\widehat{\boldsymbol{\theta}}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \int_{\mathbf{h}} p(\mathbf{h} \mid \mathbf{d}, \widehat{\boldsymbol{\theta}}) \log \underbrace{p(\mathbf{h}, \mathbf{d} \mid \boldsymbol{\theta})}_{\text{likelihood}} \, \mathrm{d}\mathbf{h}$$

- Repeat until convergence.

Stochastic EM

- A simple, stochastic version of EM that is less prone to get stuck in local maxima [5].
- Key idea: if we fill in the value of $\mathbf{h}$, the problem reverts to simple ML estimation.
- In the M-step, instead of estimating $p(\mathbf{h} \mid \mathbf{d}, \widehat{\boldsymbol{\theta}})$ we simply fill in the values of $\mathbf{h}$ by sampling.

$$\widehat{\mathbf{h}} \sim p(\mathbf{h} \mid \mathbf{d}, \widehat{\boldsymbol{\theta}})$$

  - This is often very tractable, for example using Gibbs sampling in hierarchical models.
- The E-step corresponds to ML estimation from the "completed" data set $(\widehat{\mathbf{h}}, \mathbf{d})$.

$$\widehat{\boldsymbol{\theta}}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \log p(\widehat{\mathbf{h}}, \mathbf{d} \mid \boldsymbol{\theta})$$

Pseudolikelihood estimation

- Sometimes MAP and even ML estimates are intractable. What then?
- Pseudolikelihood approximation of the ML estimate

$$\begin{aligned} p(\mathbf{d} \mid \boldsymbol{\theta}) &= p(d_1, d_2, \ldots, d_N \mid \boldsymbol{\theta}) \\ &\approx \prod_{n=1}^{N} p(d_n \mid \{d_1, \ldots, d_N\} \setminus d_n, \boldsymbol{\theta}) \end{aligned}$$

- The MPL estimate is often tractable due to conditional indepencies.
  - You often only need to condition on a small subset of $\{d_1, \ldots, d_N\} \setminus d_n$ to maximize $\boldsymbol{\theta}$.
- MPL is often used for graphical models.

Moment estimation

- A point estimate is obtained by relating the moments of a distribution (typically the mean) to its parameters.

Example: the Gamma distribution

$$\Gamma(x) \propto x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)$$

- Mean $\mu = \alpha\beta$, variance $\sigma^2 = \alpha\beta^2$
- We can estimate α and β making use of the observed mean $\widehat{\mu}$ and variance $\widehat{\sigma}$.

$$\widehat{\alpha}_{\mathrm{M}} = \frac{\widehat{\mu}^2}{\widehat{\sigma}^2} \text{ and } \widehat{\beta}_{\mathrm{M}} = \frac{\widehat{\sigma}^2}{\widehat{\mu}}$$

Summary of part I

- The Bayesian calculus offers a straightforward and systematic approach to updating a current belief in the light of new data.
- All relevant information on the parameters of interest is included in the posterior.
- The posterior can be "summarized" by various point estimates and credible intervals.
- Various point estimates can serve as useful, tractable approximations of a full Bayesian treatment.
  - Empirical Bayes, Shrinkage, ML and EM, Pseudo-ML, Moment estimation...
  - However, they will sometimes be very bad!

References part I


http://www.york.ac.uk/depts/maths/histstat/bayespic.htm


Bernardo, JM., & Smith, AF. (2009). Bayesian theory. John Wiley & Sons.


Bishop, CM. (2006). Pattern recognition and machine learning. Springer.


Boomsma, W., Mardia, KV., Taylor, CC., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008) A generative, probabilistic model of local protein structure. Proc. Natl. Acad. Sci. USA 105, 8932-8937.


Gilks, WR., Richardson, S., Spiegelhalter, DJ. Editors. (1996) Markov chain Monte Carlo in practice. Chapman & Hall.


Hamelryck, T., Mardia, KV., Ferkinghoff-Borg, J., Editors. (2012) Bayesian methods in structural bioinformatics. Book in the Springer series "Statistics for biology and health", Springer.


Lee, PM. (2012) Bayesian statistics: an introduction. John Wiley & Sons.


McGrayne, SB. (2011) The theory that would not die. Yale University Press.


Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer.