Bayesian Inference

Bayes

Laplace

Course objective

The aim of this course is to introduce the modern approach to Bayesian statistics, emphasizing the computational aspects and the differences between the classical and Bayesian approaches. The course includes Bayesian solutions to real problems.

Recommended reading

- Lee, P.M. (2004). *Bayesian Statistics: An Introduction*, (3'rd ed.), Hodder Arnold.

  A basic introduction to Bayesian statistics.

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Wiley.

  The new "bible" of Bayesian statistical theory.

- Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (2003). *Bayesian Data Analysis* (2'nd ed.), Chapman and Hall.

  An applications oriented text.

- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods* (2'nd ed.), Springer Verlag.

  All about Monte Carlo and MCMC methods.

Course evaluation

This will be based on (an exam), some applied coursework and a final applied Bayesian project from an area you are interested in.

Some recent course projects have been:

- "Bayesian learning in neural networks"
- "Bayesian arbitrage threshold analysis"
- "A Glance at Game Theory"
- "Bayesian inference for Markovian queues"
- "Estimation of objective market potential ..."

You may choose (almost) any Bayesian theme you wish.

Course outline

1. Introduction and non-Bayesian inference

- Probability and its interpretations.
- Subjective probability.
- Statistical inference.
- Classical inference. Ideas y criticisms.
- Other approaches: fiducial inference, likelihood based approaches.
- The likelihood principle.

2. Introduction to Bayesian inference: coin tossing problems

- Basic elements of Bayesian inference:
  - ◇ Bayes theorem and its interpretation
  - ◇ Prior and posterior distributions.
  - ◇ Likelihood principle.

- Coin tossing problems:
  - ◇ Results with a uniform prior.
  - ◇ Prediction.
  - ◇ Results with a beta (conjugate) prior.
  - ◇ What happens when a non-conjugate prior is used.

3. Conjugate families of distributions

- Conjugate families.
- Sufficient statistics and exponential families of distributions.
- Mixtures of conjugate distributions.
- Applications.
- Introduction to Monte Carlo sampling.

4. Gaussian models

- Inference for the normal distribution.
- The use of improper prior distributions.
- Introduction to Gibbs sampling.
- Two sample problems.
- The Behrens–Fisher problem.
- Applications.

5. Choosing a prior distribution

- Subjective prior distributions:
  - ◇ Methods for soliciting a subjective prior,
  - ◇ Problems with subjective information.

- Non informative prior distributions:
  - ◇ Insufficient reason and uniform priors,
  - ◇ Jeffreys prior distributions,
  - ◇ Maximum entropy,
  - ◇ Reference priors,
  - ◇ Problems with the use of improper and non informative prior distributions.

6. Implementation of Bayesian inference

- Numerical integration.
- Monte Carlo approaches: importance sampling and rejection sampling.
- MCMC algorithms:
  - ◇ Markov chains,
  - ◇ The Metropolis Hastings algorithm,
  - ◇ Gibbs sampling,
  - ◇ Other algorithms,
  - ◇ Implementation of Gibbs sampling via `WinBugs`.
- Applications.

7. Estimation, hypothesis testing and model choice

- Estimation as a decision problem.
- Credible intervals and the differences betwen Bayesian and classical intervals.
- Hypothesis testing:
  - ◇ Simple and composite tests.
  - ◇ Lindley's paradox.
- Bayes factors:
  - ◇ Approximations,
  - ◇ Relation to classical criteria,
  - ◇ Problems with improper priors,
  - ◇ Generalizations: Intrinsic and fractional Bayes factors.
- Application.

8. Large samples

- A Bayesian central limit theorem.
- Applications of the theorem.
- The Laplace approximation.
- Applications.

9. Regression and linear models

- Linear models.
- The two stage linear model.
- Introduction to hierarchical models and the 3 stage linear model.
- Generalized linear models.
- Applications.

10. Exchangeability and hierarchical models

- Exchangeability.
- De Finetti's theorems.
- Hierarchical models.
- Empirical Bayes approaches.
- Gibbs sampling for hierarchical models.
- Applications.

11. Time series and dynamic linear models

- Dynamic linear models:
  - ◇ The closed, constant DLM,
  - ◇ Relations with classical approaches,
  - ◇ Modelling trend and seasonality.

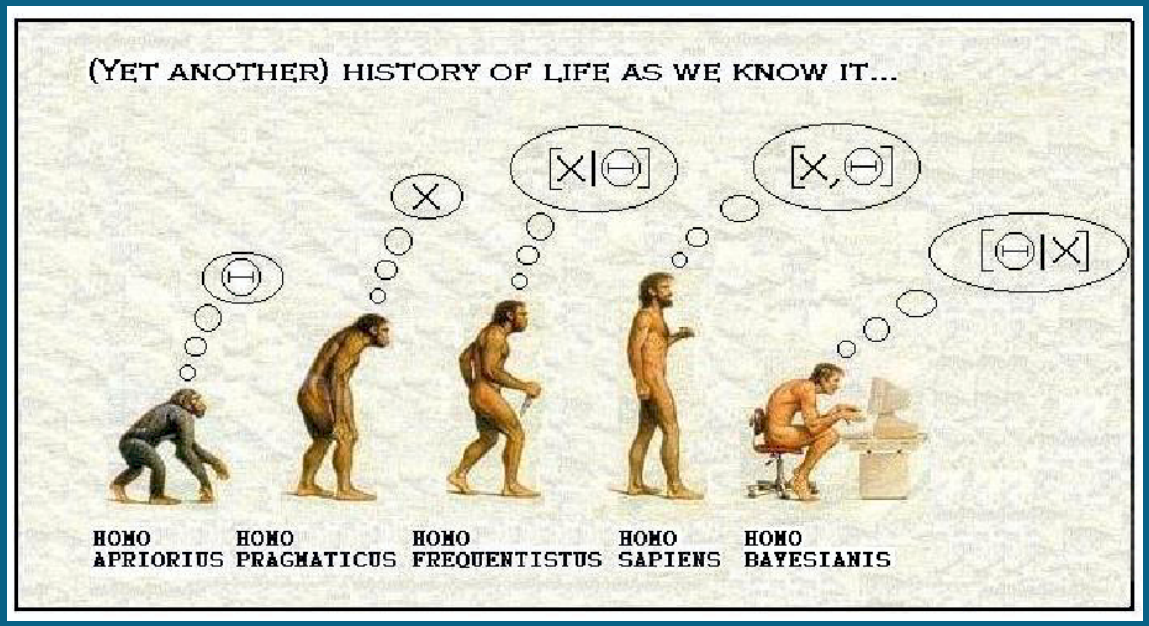
- Other approaches to time series.

12. Other topics

Selected from:

- Robustness and sensitivity analysis.
- Bayesian nonparametric methods.
- Graphical models, belief nets and Bayes linear models.
- Decision analysis.

1. Introduction and non-Bayesian inference



Objective

Introduce the different *objective* and *subjective* interpretations of probability. Examine the various non-Bayesian treatments of statistical inference and comment on their associated problems.

Recommended reading

- Hájek, A. (2003). Interpretations of Probability. In *Stanford Encyclopedia of Philosophy*.

  http://plato.stanford.edu/entries/probability-interpret/

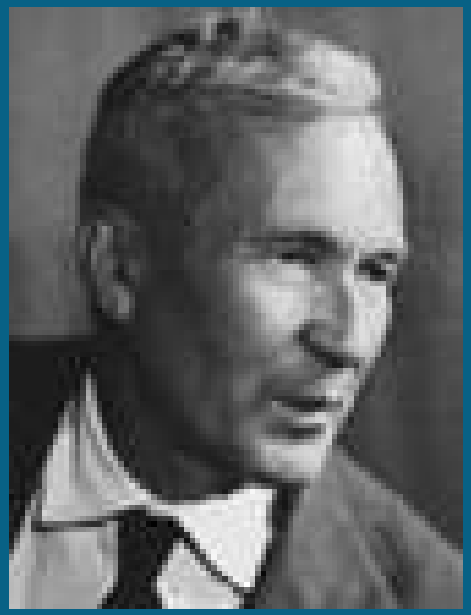
- The Wikipedia has a nice page on different interpretations of probability.

  http://en.wikipedia.org/wiki/Probability_interpretations

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Chapter 2.

  http://www.uv.es/bernardo/BT2.pdf

Probability



Kolmogorov

Probability theory developed from studies of games of chance by Fermat and Pascal and may be thought of as the study of randomness. It was put on a firm mathematical basis by Kolmogorov (1933).

The Kolmogorov axioms

For a *random experiment* with *sample space* $\mathbf{\Omega}$, then a probability measure P is a function such that

1. for any event $A \in \mathbf{\Omega}$, $P(A) \geq 0$.

2. $P(\mathbf{\Omega}) = 1$.

3. $P\left(\cup_{j \in J} A_j\right) = \sum_{j \in J} P(A_j)$ if $\{A_j : j \in J\}$ is a countable set of incompatible events.

The laws of probability can be derived as consequences of these axioms. However, this is a purely mathematical theory and does not provide a useful practical interpretation of probability.

Interpretations of probability

There are various different ways of interpreting probability.

- The classical interpretation.
- Logical probability.
- Frequentist probability.
- Propensities.
- Subjective probability.

For a full review, see e.g. Gillies (2000).

Classical probability

Bernoulli

This derives from the ideas of Jakob Bernoulli (1713) contained in the *principle of insufficient reason* (or *principle of indifference*) developed by Laplace (1814) which can be used to provide a way of assigning epistemic or subjective probabilities.

The principle of insufficient reason

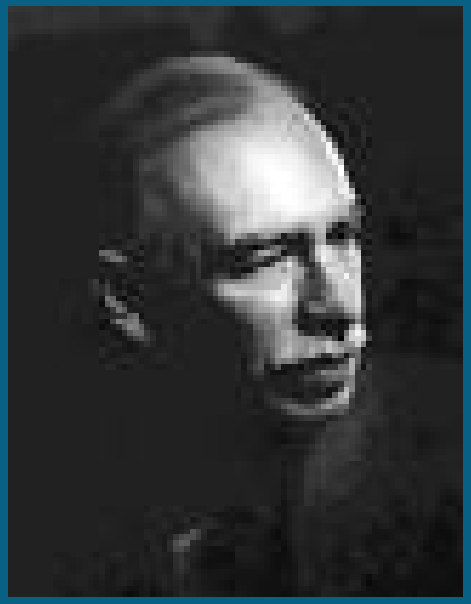
If we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way.

Thus the probability of an event is the coefficient between the number of favourable cases and the total number of possible cases.

This is a very limited definition and cannot be easily applied in infinite dimensional or continuous sample spaces.

http://en.wikipedia.org/wiki/Principle_of_indifference

Logical probability



Keynes

Carnap

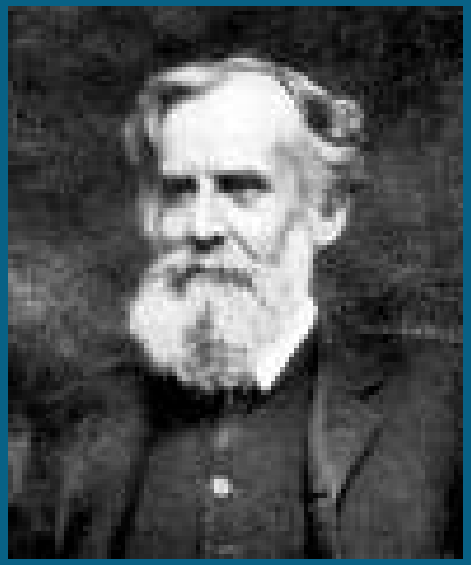
This approach, which extends the classical concept of probability, was developed by Keynes (1921) and Carnap (1950). The probability of a proposition H given evidence E is interpreted as the (unique) degree to which E logically entails H.

Logical probabilities are constructed using the idea of formal languages as follows:

- Consider a language, L, with predicates $H_1, H_2, \ldots$ and a finite number of constants $e_1, e_2, \ldots, e_n$.
- Define a probability measure $P(\cdot)$ over sentences in L in a way that only takes into account their syntactic structure.
- Then use the standard probability ratio formula to create conditional probabilities over pairs of sentences in L.

Unfortunately, as noted by Bernardo and Smith (1994), "the logical view of probability is entirely lacking in operational content". Such probabilities are assumed to exist and depend on the formal language in which they are defined.

Frequentist probability



Venn

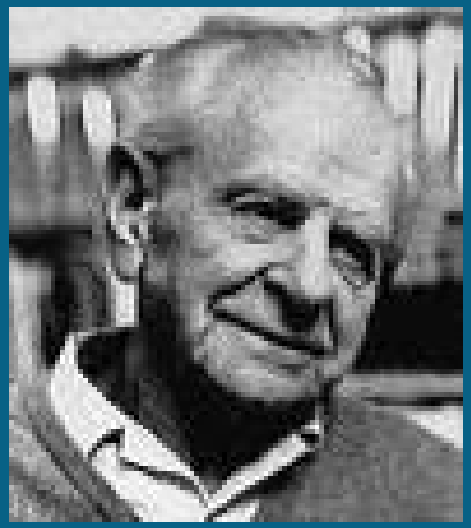
Von Mises

The idea comes from Venn (1876) and was expounded by von Mises (1919).

Given a repeatable experiment, the probability of an event is the limit of the proportion of times that the event will occur when the number of repetitions of the experiment tends to infinity.

This is a restricted definition of probability. It is impossible to assign probabilities in non repeatable experiments.

Propensities



Popper

This theory was developed by Popper (1957).

Probability is an innate disposition or propensity for things to happen. In particular, long run propensities seem to coincide with the frequentist definition of probability whereas it is not clear what individual propensities are, or whether they obey the probability calculus.

Subjective probability

Ramsey

The subjective concept of probability is as *degrees of belief*. A first attempt to formalize this was made by Ramsey (1926).

In reality, most people are irrational in that their own degrees of belief do not satisfy the probability axioms, see e.g. Kahneman et al (1982) and chapter 5 of this course. Thus, in order to formalize the definition of subjective probability, it is important to only consider *rational agents*, i.e. agents whose beliefs are logically consistent.

Consistent degrees of belief are probabilities

Cox (1946) formalized the conditions for consistent reasoning by assuming defining the necessary logical conditions. Firstly, relative beliefs in the truths of different propositions are assumed to be transitive, i.e. if we believe that $A \succeq B$ and $B \succeq C$, for three events A, B, C, then we must believe that $A \succeq C$, where $A \succeq B$ means A is at least as likely as B to occur. This assumption implies that we can represent degrees of belief by numbers, where the higher the value, the higher the degree of belief.

Secondly, it is assumed that if we specify how much we believe an event A to be true, then we also implicitly specify how much we believe it to be false.

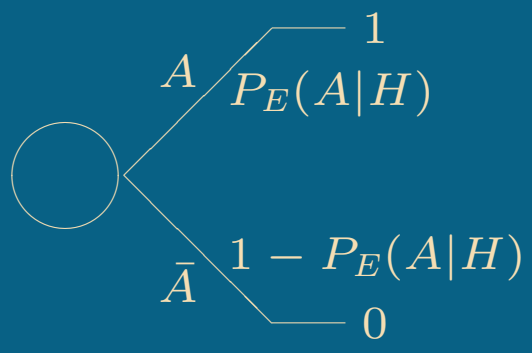
Finally it is assumed that if we specify how much we believe A to be true and how much we believe B to be true, given that A is true, then we are implicitly implying our degree of belief that both A and B are true.

Given these axioms, Cox was able to demonstrate that logical consistency could only be assured if the numbers used to represent degrees of belief are probabilities.

Defining subjective probabilities

Cox's method is non-constructive and does not give a method of defining a probability for a given event. There are various ways of doing this, usually based around the ideas of betting.

De Finetti (1937) defines la probability, $P_E(A|H)$, of an event A for an agent E with information H to be the maximum quantity, or fair price, that E would pay for a ticket in a lottery where they will obtain a (small) unit prize if and only if A occurs.



The expected gain in this lottery, is equal to

$$P_E(A|H) \times 1 + (1 - P_E(A|H)) \times 0 = P_E(A|H).$$

A *Dutch book* is defined to be a series of bets, each acceptable to the agent, which collectively imply that the agent is certain to lose money, however the world turns out.

It can now be shown that in order to avoid a Dutch book, then the agent's probabilities must satisfy the usual probability calculus.

Note that the definition proposed by De Finetti depends on the assumption that the *utility* of the agent for money is linear. A more general method of defining probabilities and utilities is provided by Savage (1954). See also O' Hagan (1988) chapters 1 to 3 for a definition based on betting and odds.

Statistical Inference

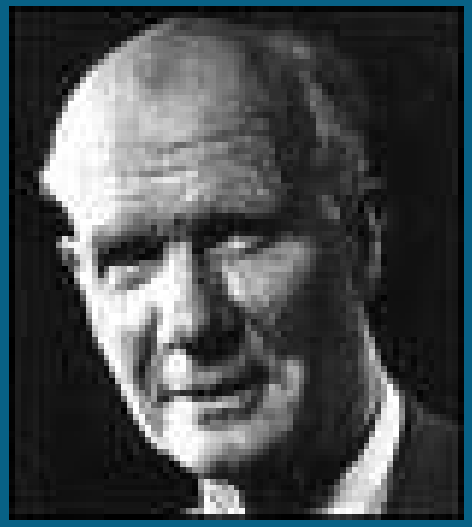
A number of different approaches to statistical inference have been developed based on both the frequentist and subjective concepts of probability.

- Classical (frequentist) inference.
- Likelihood based approaches.
- Fiducial statistics and related methods.
- Bayesian inference.

For full comparisons of the different approaches see e.g. Barnett (1999).

Classical inference

Neyman



Pearson

This approach was developed from the ideas of Neyman and Pearson (1933) and Fisher (1925).

Characteristics of classical inference

- Frequentist interpretation of probability.
- Inference is based on the likelihood function $l(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$.
- We can only quantify (a priori) the uncertainty about $\mathbf{X}$. $\boldsymbol{\theta}$ is fixed.
- Inferential procedures are based on asymptotic performance:
  - ◇ An estimator $\mathbf{t} = \mathbf{t}(\mathbf{X})$ is defined.
  - ◇ The plausibility of a given value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is measured by the density

$$l(\boldsymbol{\theta}_0|\mathbf{x}) \propto f(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}_0),$$

    (assuming that $t(\cdot)$ is a sufficient statistic).
  - ◇ If $\mathbf{t}$ does not lie in the tail of $f(\mathbf{t}|\boldsymbol{\theta}_0)$ then $\boldsymbol{\theta}_0$ is a plausible value of $\boldsymbol{\theta}$.

Classical estimation and hypothesis tests

Classical point estimation is based on choosing estimators with good asymptotic properties (unbiasedness, minimum variance, efficiency etc.)

There are a number of possible methods of choosing an estimator, e.g. method of moments, maximum likelihood, etc. Note that in particular, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined so that $l(\hat{\boldsymbol{\theta}}|\mathbf{x}) > l(\boldsymbol{\theta}|\mathbf{x})$ for all $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$. The MLE is asymptotically unbiased, efficient etc.

For interval estimation, an interval $(l(\mathbf{x}),\ u(\mathbf{x}))$ is chosen so that

$$P(l(\mathbf{X}) < \boldsymbol{\theta} < u(\mathbf{X})|\boldsymbol{\theta}) = 1 - \alpha$$

for some fixed probability level α.

Finally, hypothesis testing is based on rejecting $\boldsymbol{\theta}_0$ at level α if

$$P(t(\mathbf{X}) > t(\mathbf{x})|\boldsymbol{\theta}_0) < \alpha.$$

Principles justifying classical inference

The principle of sufficiency

Definition 1
A statistic $\mathbf{t} = \mathbf{t}(\mathbf{x})$ is said to be sufficient for $\boldsymbol{\theta}$ if

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = f(\mathbf{x}|\mathbf{t}).$$

The *sufficiency principle* (Fisher 1922) is as follows.

If a sufficient statistic, $\mathbf{t}$, exists, then for any two samples $\mathbf{x}_1$, $\mathbf{x}_2$ of the same size such that $\mathbf{t}(\mathbf{x}_1) = \mathbf{t}(\mathbf{x}_2)$ then the conclusions given $\mathbf{x}_1$ and $\mathbf{x}_2$ should be the same.

All standard methods of inference satisfy this principle.

The Fisher-Neyman factorization theorem

This gives a useful characterization of a sufficient statistic which we shall use in Chapter 3.

Theorem 1
A statistic $\mathbf{t}$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions g and h such that

$$l(\boldsymbol{\theta}|\mathbf{x}) = g(\mathbf{t}, \boldsymbol{\theta})h(\mathbf{x}).$$

Proof See e.g. Bernardo and Smith (1994). ∎

The principle of repeated sampling

The inference that we draw from $\mathbf{x}$ should be based on an analysis of how the conclusions change with variations in the data samples, which would be obtained through hypothetical repetitions, under exactly the same conditions, of the experiment which generated the data x in the first place.

This principle follows directly from the frequentist definition of probability and is much more controversial. It implies that measures of uncertainty are just hypothetical asymptotic frequencies. Thus, it is impossible to measure the uncertainty about $\boldsymbol{\theta}$ *a posteriori*, given $\mathbf{x}$.

Criticisms of classical inference

Global criticisms

As noted earlier, the frequentist definition of probability restricts the number of problems which can be reasonably analyzed. Furthermore, classical inference appears to be a form of *cookbook* containing many seemingly *ad-hoc* procedures.

Specific criticisms

We can also criticize many specific aspects of the classical approach.

Firstly, there are often problems with estimation. An optimal method of choosing an estimator is not generally available and some typically used approaches may provide very bad estimators.

One possibility is to attempt to use an unbiased estimator.

Example 1

Let X be an observation from a Poisson distribution: $\mathcal{P}(\lambda)$ and assume that we wish to estimate $\phi = e^{-2\lambda}$. Then only one unbiased estimator of ϕ exists and takes the value $\tilde{\phi} = (-1)^X = \pm 1$. However $0 < \phi \leq 1$ for all λ.

One possibility is to attempt to use an unbiased estimator.

Example 1
Let X be an observation from a Poisson distribution: $\mathcal{P}(\lambda)$ and assume that we wish to estimate $\phi = e^{-2\lambda}$. Then only one unbiased estimator of ϕ exists and takes the value $\tilde{\phi} = (-1)^X = \pm 1$. However $0 < \phi \leq 1$ for all λ.

An alternative is to use the method of moments. However, this approach is not always available.

Example 2
Suppose that we wish to estimate the location parameter θ of a Cauchy distribution. Then no simple method of moments estimator exists.

Also, method of moments of estimators, when they do exist do not have the optimality properties of e.g. maximum likelihood estimators.

It is more common to use maximum likelihood estimators. These are justified asymptotically but can have very bad properties in small samples.

Example 3
Let $X \sim \mathcal{DU}[1, \theta]$. Then given a sample of size 1, the MLE of θ is $\hat{\theta} = X$. The bias of the MLE is

$$E[X] - \theta = \frac{\theta + 1}{2} - \theta = \frac{1 - \theta}{2},$$

which can be enormous if θ is large.

Whatever the sample size, the MLE is always underestimates the true value in this experiment.

Example 4

Suppose that $\mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right)$ where $\dim(\boldsymbol{\theta}) = n$, and that we wish to estimate $\boldsymbol{\theta}$ given $\mathbf{Y} = \mathbf{y}$. Then clearly, the maximum likelihood estimator is $\hat{\boldsymbol{\theta}} = \mathbf{y}$.

Example 4

Suppose that $\mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}\right)$ where $\dim(\boldsymbol{\theta}) = n$, and that we wish to estimate $\boldsymbol{\theta}$ given $\mathbf{Y} = \mathbf{y}$. Then clearly, the maximum likelihood estimator is $\hat{\boldsymbol{\theta}} = \mathbf{y}$.

However, if $n \geq 3$ then this estimator is inadmissible as the James-Stein estimator $\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}$ has lower mean squared error.

We will return to this example in Chapter 10.

A more important problem is interpretation.

Example 5

Suppose that we carry out an experiment and find that a 95% confidence interval for θ based on the sample data is equal to $(1, 3)$. How do we interpret this?

A more important problem is interpretation.

Example 5
Suppose that we carry out an experiment and find that a 95% confidence interval for θ based on the sample data is equal to $(1, 3)$. How do we interpret this?

This means that if we repeated the same experiment and procedure by which we constructed the interval many times, 95% of the constructed intervals would contain the true value of θ.

A more important problem is interpretation.

Example 5
Suppose that we carry out an experiment and find that a 95% confidence interval for θ based on the sample data is equal to $(1, 3)$. How do we interpret this?

This means that if we repeated the same experiment and procedure by which we constructed the interval many times, 95% of the constructed intervals would contain the true value of θ.

It does not mean that the probability that θ lies in the interval $(1, 3)$ is 95%.

There are often problems in dealing with nuisance parameters in classical inference.

Example 6
Let $Y_{i,j} \sim \mathcal{N}\left(\phi_i, \sigma^2\right)$ for $i = 1, \ldots, n$ and $j = 1, 2$.

Suppose that the parameter of interest is the variance, σ^2 and that $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are nuisance parameters.

The likelihood is $l\left(\sigma^2, \boldsymbol{\phi}|\mathbf{y}\right) \propto$

$$\sigma^{-2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{i,1}-\phi_i)^2 + (y_{i,2}-\phi_i)^2\right).$$

and now, the most natural way of estimating σ^2 is to maximize the *profile likelihood*. This is calculated as follows.

Supposing σ^2 known, we first maximize the likelihood with respect to ϕ.

$$\begin{aligned} l_P(\sigma^2|\mathbf{y}) &= \sup_{\boldsymbol{\phi}} l\left(\sigma^2, \boldsymbol{\phi}|\mathbf{y}\right) \\ &= \sigma^{-2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{i,1}-\bar{y}_i)^2+(y_{i,2}-\bar{y}_i)^2\right) \\ &= \sigma^{-2n} \exp\left(-\frac{1}{4\sigma^2}\sum_{i=1}^{n}(y_{i,1}-y_{i,2})^2\right) \end{aligned}$$

Secondly, we maximize the profile likelihood with respect to σ^2, which implies that

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}(y_{i,1}-y_{i,2})^2.$$

However,

$$E\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{2}$$

for any value of n and this estimator is inconsistent.

It is also not entirely clear how predictions should be carried out.

Example 7

Let $X|\theta \sim f(\cdot|\theta)$. Then typically, the predictive distribution is estimated, given the sample data, by substituting θ by its MLE, leading to the estimated predictive density $f(x|\hat{\theta})$. However this procedure clearly underestimates the predictive uncertainty due to the fact that θ is unknown.

Likelihood based inference

Barnard

This approach is based totally on the likelihood function and derives from Barnard (1949) and Barnard et al (1962). Firstly, the likelihood principle is assumed.

The likelihood principle

This principle is originally due to Barnard (1949). Edwards (1992) defines the likelihood principle as follows.

Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information.

This is equivalent to supposing that if we have experiments, E_i of the same size with sample data $\mathbf{x}_i$ and sampling distributions $f_i(\cdot|\boldsymbol{\theta})$ for $i = 1, 2$, then the two experiments provide the same evidence (EV) and hence the same inference about $\boldsymbol{\theta}$ if $f_1(\mathbf{x}_1|\boldsymbol{\theta}, E_1) \propto f_2(\mathbf{x}_2|\boldsymbol{\theta}, E_2)$, that is

$$l(\boldsymbol{\theta}|E_2, \mathbf{x}_2) = cl(\boldsymbol{\theta}|E_1, \mathbf{x}_1) \quad \text{for some } c \;\Rightarrow EV[E_1, \mathbf{x}_1] = EV[E_2, \mathbf{x}_2].$$

A second assumption of likelihood inference is that the likelihood ratio for $l(\boldsymbol{\theta}_1|\mathbf{x})/l(\boldsymbol{\theta}_2|\mathbf{x})$, for any two values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is a measure of the evidence supplied by the data in favour of $\boldsymbol{\theta}_1$ relative to $\boldsymbol{\theta}_2$.

This is much harder to apply in practice, in particular in the presence of nuisance parameters. For example, how can we marginalize the likelihood function?

Stopping rules

Stopping rules are often used in classical statistics (e.g. clinical trials) to make it possible to stop early if the results are sufficiently favourable or unfavourable.

Frequentist statisticians must choose the stopping rule before the experiment begins and must stick to it exactly (otherwise the good frequentist properties of the test are lost).

Example 8
Assume we are interested in testing whether a coin is biased in favour of heads. Then we may consider various different stopping rules, e.g.

- toss the coin a fixed number, n times.
- toss the coin until the first tail is seen.
- toss the coin until the r'th tail is seen.

The stopping rule principle

The stopping rule principle is the following.

In a sequential experiment, the evidence provided by the experiment about the value of the unknown parameters θ should not depend on the stopping rule.

This is clearly a consequence of the likelihood principle as the likelihood function is independent of the stopping rule. However ...

Classical hypothesis tests do not satisfy the stopping rule principle

Example 9

Suppose that $\theta = P(\text{head})$. We wish to test $H_0 : \theta = 1/2$ against the alternative $H_1 : \theta > 1/2$ at a 5% significance level.

Suppose that we observe 9 heads and 3 tails. This information is not sufficient for us to write down the likelihood function. We need to know the sampling scheme or stopping rule.

Suppose that we fix the number of tosses of the coin to be 12. Then

$$X = \#\text{ heads}|\theta \sim \mathcal{BI}(12, \theta) \quad \text{and} \quad l(\theta|x = 9) = \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3.$$

Thus, the p-value is

$$p_1 = \frac{1}{2}^{12} \left[\left(\begin{array}{c} 12 \\ 9 \end{array} \right) + \cdots + \left(\begin{array}{c} 12 \\ 12 \end{array} \right) \right] \approx .075$$

and we do not reject the null hypothesis.

Suppose now that we decided to calculate the number of heads X until the third tail occurs. Thus, $X|\theta \sim \mathcal{NB}(3,\theta)$ and

$$l(\theta|x) = \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9(1-\theta)^3 \quad \text{and the p-value is}$$

$$\begin{aligned} p_2 &= \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9(1-\theta)^3 + \left(\begin{array}{c} 12 \\ 10 \end{array} \right) \theta^{10}(1-\theta)^3 + \cdots \\ &= .0325 \quad \text{and we reject the null hypothesis.} \end{aligned}$$

The reason is that in order to carry out a hypothesis test, then we must specify the sample space or stopping rule. This is different in the two cases that we have seen:

1. $\Omega = \{(u, d) : u + d = 12\}$

2. $\Omega = \{(u, d) : d = 3\}$

The conditionality principle

Suppose that we have the possibility of carrying out two experiments E_1 and E_2 in order to make inference about $\boldsymbol{\theta}$ and that we choose the experiment to carry out by tossing an unbiased coin. Then our inference for $\boldsymbol{\theta}$ should only depend on the selected experiment.

To formalize this, note that the experiment E_i may be characterized as $E_i = (\mathbf{X}_i, \boldsymbol{\theta}, f_i)$ which means that in this experiment, the variable $\mathbf{X}_i$ is generated from $f_i(\mathbf{X}_i|\boldsymbol{\theta})$.

Now define the composite experiment $E^\star$ which consists of generating a random variable K where $P(K = 1) = P(K = 2) = \frac{1}{2}$ and then performing experiment K so that $E^\star = (\underbrace{(K, \mathbf{X}_K)}_{\mathbf{x}}, \boldsymbol{\theta}, \frac{1}{2} f_K(\mathbf{X}_K))$. Then conditionality implies that $EV[E^\star, \mathbf{x}] = \begin{cases} EV[E_1, \mathbf{x}_1] & \text{if } K = 1 \text{ so } \mathbf{x} = (1, \mathbf{x}_1) \\ EV[E_2, \mathbf{x}_2] & \text{if } K = 2 \text{ so } \mathbf{x} = (1, \mathbf{x}_2) \end{cases}$

Birnbaum (1962) demonstrates the following theorem which relates the sufficiency, likelihood and conditionality principles.

Birnbaum (1962) demonstrates the following theorem which relates the sufficiency, likelihood and conditionality principles.

Theorem 2

The likelihood principle is equivalent to the sufficiency principle and the conditionality principle.

We shall demonstrate that sufficiency plus conditionality $\Rightarrow$ likelihood. See Birnbaum (1962) for a full proof.

Proof of the theorem

Proof Let E_1, E_2 and $\mathbf{x}_1$ and $\mathbf{x}_2$ be the two experiments and data samples figuring in the statement of the likelihood principle so that

$$l(\boldsymbol{\theta}|E_1, \mathbf{x}_1) = cl(\boldsymbol{\theta}|E_2, \mathbf{x}_2)$$

for some constant c and define the composite experiment $E^\star = E_1$ as in the conditionality principle.

Define the statistic $\mathbf{t}$ by

$$\mathbf{t} = \mathbf{t}(\mathbf{x}) = \begin{cases} (1, \mathbf{x}_1) & \text{if } K = 2 \\ \mathbf{x} & \text{otherwise} \end{cases}$$

Now note that if $\mathbf{t} \neq (1, \mathbf{x}_1)$ then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \mathbf{t} = \mathbf{t}(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

whereas if $\mathbf{t} = \mathbf{x} = (1, \mathbf{x}_1)$, then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \frac{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta})}{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta}) + \frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})} = \frac{c}{1+c}$$

and if $\mathbf{t} = (1, \mathbf{x}_1)$ but $\mathbf{x} = (2, \mathbf{x}_2)$ then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \frac{\frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})}{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta}) + \frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})} = \frac{1}{1+c}$$

so $f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = f(\mathbf{x}|\mathbf{t})$ and $\mathbf{t}$ is a sufficient statistic.

It follows from the sufficiency principle that

$$EV[E^\star, (1, \mathbf{x}_1)] = EV[E^\star, (2, \mathbf{x}_2)]$$

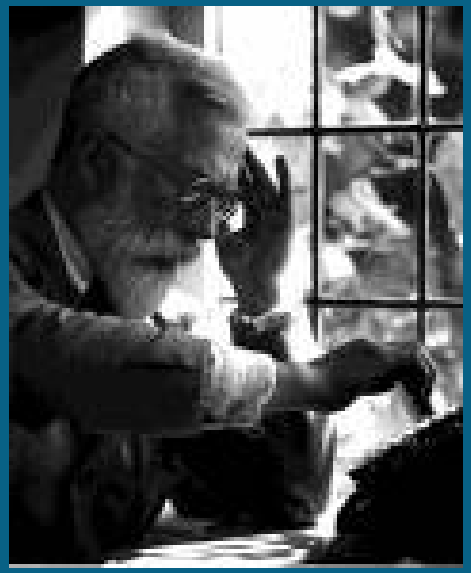
and then the conditionality principle ensures that

$$EV[E^\star, \mathbf{x}_1] = EV[E_1, (1, \mathbf{x}_i)] = EV[E^\star, (2, \mathbf{x}_2)] = EV[E_2, \mathbf{x}_2]$$

which is the likelihood principle. ∎

In the same way, it can be demonstrated that likelihood + sufficiency $\Rightarrow$ conditionality or that likelihood + conditionality $\Rightarrow$ sufficiency.

Fiducial inference and related methods



Fisher

Fiducial inference has the objective of defining a posterior measure of uncertainty for θ without the necessity of defining a prior measure. This approach was introduced by Fisher (1930).

Example 10

Let $X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we wish to carry out inference for μ.

We know that $T = \frac{(\bar{X}-\mu)\sqrt{n}}{S} \sim t_{n-1}$. Then for any t, $P(T > t) = p(t)$ where $p(t)$ is known. Fisher's idea is to write

$$\begin{aligned} p(t) &= P(T > t) \\ &= P\left(\frac{(\bar{X}-\mu)\sqrt{n}}{S} > t\right) \\ &= P\left(\mu < \bar{X} - \frac{St}{\sqrt{n}}\right) \end{aligned}$$

and then define $p(t) = P\left(\mu < \bar{x} - \frac{st}{\sqrt{n}}\right)$ to be the fiducial probability that μ is less than $\bar{x} - \frac{st}{\sqrt{n}}$.

Problems with the fiducial approach

- The probability measure is transferred from the sample space to the parameter space. What is the justification for this?
- What happens if no pivotal statistic exists?
- It is unclear how to apply the fiducial approach in multidimensional problems.

In many cases, fiducial probability intervals coincide with Bayesian credible intervals given specific non informative prior distributions. In particular, structural inference, see Fraser (1968) corresponds to Bayesian inference using so called Haar prior distributions. However we will see that the Bayesian justification for such intervals is more coherent.

References

Barnard, G.A. (1949). Statistical Inference, *Journal of the Royal Statistical Society. Series B*, *11*, 115–149.

Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1962) Likelihood Inference and Time Series. *Journal of the Royal Statistical Society, Series A*, **125**, 321–372.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. New York: Wiley.

Barnett, V. (1999). *Comparative Statistical Inference* (3rd ed.). Chichester: Wiley

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis*. Basel: Thurneysen Brothers.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, **57**, 269–306.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Cox, R. (1946). Probability, Frequency, and Reasonable Expectation, *Am. Jour. Phys.*, **14**, 1–13.

Edwards, A.W.F. (1992). *Likelihood* (2nd ed.) Cambridge: University Press.

Finetti, R. de (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1–68.

Fisher, R.A. (1922). On the Mathematical Foundations Of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309-68.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, **26**, 528–535.

Fraser, D.A.S. (1968). *The Structure of Inference*. New York: Wiley.

Gillies, D. (2000). *Philosophical theories of probability*. Routledge.

Hájek, A. (2003). Interpretations of Probability. In *Stanford Encyclopedia of Philosophy*.

Kahneman, D., Slovic, P. and Tversky, A. (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: University Press.

Keynes, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Laplace, P.S. (1814). *Théorie Analytique des probabilités*. Paris: Courcier Imprimeur.

Mises, R. von (1919). *Wahrscheinlichkeit, Statistik und Wahrheit*. Vienna, Springer.

Neyman, J. and Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London. Series A*, **231**, 289–337.

O' Hagan, A. (1988). *Probability: Methods and Measurement*. London: Chapman and Hall.

Pederson, J.G. (1978). Fiducial Inference. *International Statistical Review*, **46**, 147-170.

Popper, K. (1957). The Propensity Interpretation of the Calculus of Probability and of the Quantum Theory. In *Observation and Interpretation.* Buttersworth Scientific Publications.

Ramsey, F.P. (1926). Truth and Probability. In *Foundations of Mathematics and other Essays*. Routledge and Kegan.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

Venn, J. (1876). *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability.*

2. Introduction to Bayesian inference

de Finetti

The Bayesian approach comes originally from the work of Bayes and Laplace and much of the modern theory comes from de Finetti in the 1930's.

Recommended reading

- Este artículo de José Bernardo es una buena introducción a la inferencia bayesiana.

- Lindley (1983) is a useful article to read.

Characteristics of Bayesian inference

Firstly, Bayesian inference depends directly on the subjective definition of probability.

We can all have our own probabilities for a given event:

$$P(\text{head}),\ P(\text{rain tomorrow}),\ P(\text{Mike was born in 1962}).$$

Our probabilities may be different as they are our own measures of the likelihood of given events.

Secondly, given a sample $\mathbf{x}$, and a prior distribution $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, we can update our beliefs using $Bayes\ theorem$:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &= \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\mathbf{x})} \\ &\propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta}) \end{aligned}$$

Bayesian inference satisfies the likelihood principle

Proof Suppose that $l(\boldsymbol{\theta}|\mathbf{x}_1) \propto l(\boldsymbol{\theta}|\mathbf{x}_2)$ and assume a prior distribution $p(\boldsymbol{\theta})$. Then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}_1) &\propto p(\boldsymbol{\theta}) l(\boldsymbol{\theta}|\mathbf{x}_1) \\ &\propto p(\boldsymbol{\theta}) l(\boldsymbol{\theta}|\mathbf{x}_2) \\ &\propto p(\boldsymbol{\theta}|\mathbf{x}_2) \\ &= p(\boldsymbol{\theta}|\mathbf{x}_2). \end{aligned}$$

∎

In the above, we are assuming that the prior distribution is subjective, i.e. that it is not just chosen via formal rules, e.g. always uniform. In this case, the likelihood principle can be violated. See chapter 5.

Estimation and credible intervals

For a Bayesian, estimation is treated as a decision problem. In a given situation, we should elect an estimator in order to minimize the loss that we expect to incur. *Utility theory* can be used to choose an optimal estimator. See chapter 7.

A 95% credible interval for θ is an interval $[a, b]$ such that our probability that θ lies in $[a, b]$ is 95%. See chapter 5 for a formal definition.

Nuisance parameters and prediction

There are no (theoretical) problems in dealing with with nuisance parameters.

If $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2$ are nuisance parameters, then we can write the joint density as $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2)$ and therefore,

$$p(\boldsymbol{\theta}_1|\mathbf{x}) = \int p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x})p(\boldsymbol{\theta}_2|\mathbf{x})\, d\boldsymbol{\theta}_2.$$

Note however that if $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x})$ varies a lot given different values of $\boldsymbol{\theta}_2$, then this sensitivity should be taken into account. See Box and Tiao (1992), Section 1.6.

Prediction is also straightforward. If Y is a new observation, then the predictive distribution of Y is

$$f(y|\mathbf{x}) = \int f(y|(\mathbf{x}), \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}$$

Bayesian analysis of Example 9

Example 11

Suppose that our prior beliefs about θ are represented by a uniform distribution $\theta \sim \mathcal{U}(0, 1)$.

The uniform distribution is an example of a beta distribution,

$$p(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \phi^{\alpha-1}(1-\phi)^{\beta-1} \quad \text{for } 0 < \phi < 1$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function. Setting $\alpha = \beta = 1$ gives the uniform distribution. This is not a very realistic representation of typical prior knowledge. It would be more appropriate to use a symmetric beta distribution, e.g. $\mathcal{B}(5, 5)$.

We can now calculate the posterior distribution via Bayes theorem.

Calculation of the posterior distribution

From Bayes theorem, the posterior distribution is

$$\begin{aligned} p(\theta|x) &\propto 1 \times \binom{12}{9} \theta^9(1-\theta)^3 \\ &\propto \theta^9(1-\theta)^3 \propto \theta^{10-1}(1-\theta)^{4-1} \\ &= \frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} \end{aligned}$$

which implies that $\theta|\mathbf{x} \sim \mathcal{B}(10,4)$.

It can now be demonstrated that $P(\theta \leq 1/2|\mathbf{x}) \approx .046$ and we might choose to reject the hypothesis that $\theta \leq 0.5$. Note however that this does not constitute a formal hypothesis test. These will be analyzed in chapter 7.

The posterior mean as a weighted average

From the properties of the beta distribution, we know that if $\phi \sim \mathcal{B}(\alpha, \beta)$, then $E[\phi] = \frac{\alpha}{\alpha+\beta}$.

Thus, in our case, we have

$$E[\theta|x] = \frac{10}{10+4} = \frac{5}{7} \quad \text{and moreover,}$$

$$\frac{5}{7} = \frac{1}{7} \times \frac{1}{2} + \frac{6}{7} \times \frac{9}{12}$$

which implies that

$$E[\theta|x] = \frac{1}{7}E[\theta] + \frac{6}{7}\hat{\theta}$$

where $E[\theta] = 1/(1+1) = 1/2$ is the prior mean and $\hat{\theta} = 9/12$ is the MLE of θ.

Thus, the posterior mean is a weighted average of the prior mean and the MLE.

Interpretation

One interpretation of this result is that the data have six times the weight of the prior distribution in determining the posterior distribution.

Equally, we might assume that the information represented in the prior distribution is equivalent to the information contained in an experiment where a coin is tossed twice and one head and one tail are observed. This interpretation will be formalized in the following chapter.

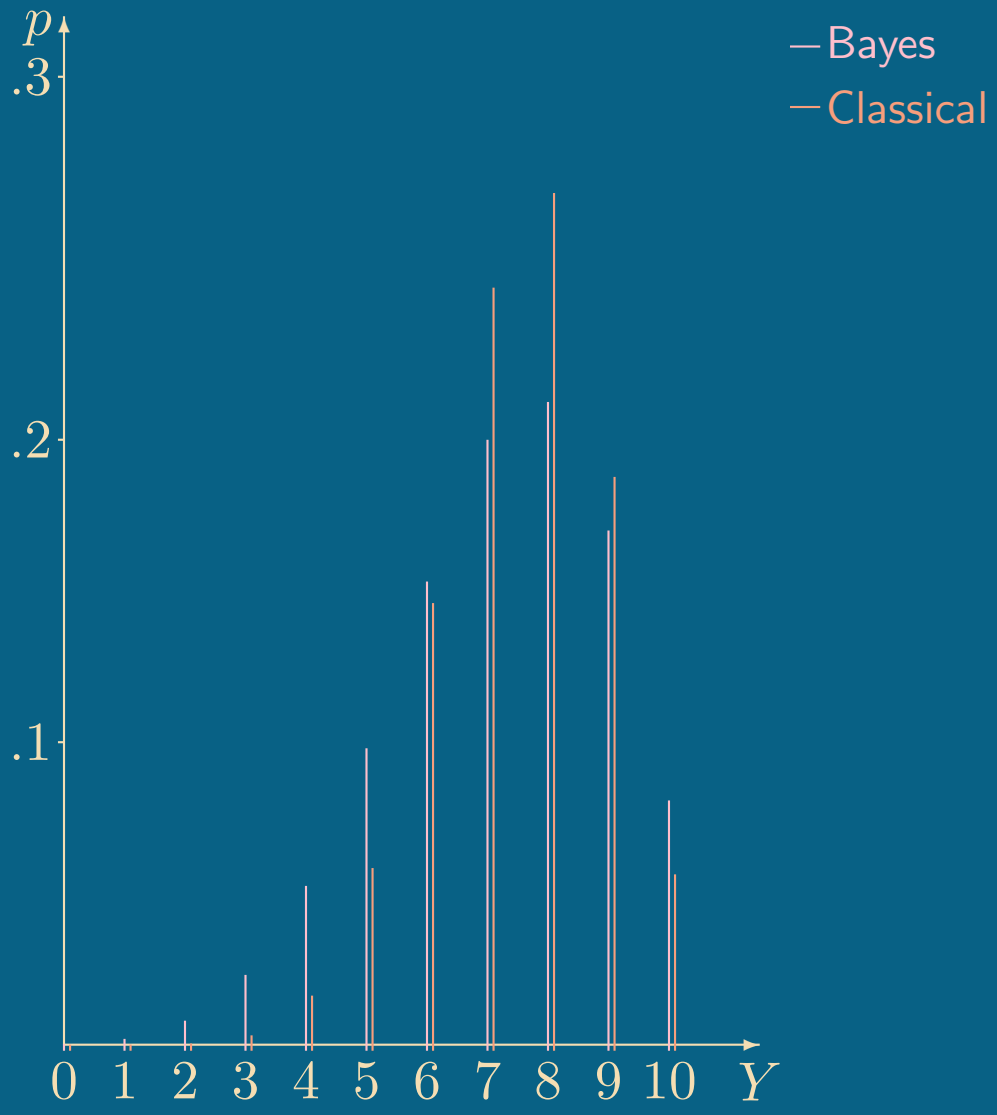
Prediction

Suppose that we wish to predict the number of heads, Y, in ten further tosses of the same coin. Thus, we have $Y|\theta \sim \mathcal{BI}(10, \theta)$ and therefore,

$$
\begin{aligned}
f(y|\mathbf{x}) &= \int f(y|\mathbf{x}, \theta) p(\theta|\mathbf{x})\, d\theta = \int f(y|\theta) p(\theta|\mathbf{x})\, d\theta \\
&= \int_0^1 \binom{10}{y} \theta^y (1-\theta)^{10-y} \times \frac{1}{B(10,4)} \theta^{10-1} (1-\theta)^{4-1}\, d\theta \\
&= \binom{10}{y} \frac{1}{B(10,4)} \times \int_0^1 \theta^{10+y-1} (1-\theta)^{14-y-1}\, d\theta \\
&= \binom{10}{y} \frac{B(10+y, 14-y)}{B(10,4)} \quad \text{for } y = 0, 1, \ldots, 10
\end{aligned}
$$

which is the so called *beta-binomial distribution.* The following diagram illustrates the predictive probability distribution of Y and the binomial predictive distribution ($\mathcal{BI}(10, .75)$) derived from substituting the MLE, $\hat{p} = 0.75$, for p.

The predictive distribution



The predictive mean and variance

We can calculate the predictive mean of $Y|\mathbf{x}$ without having to evaluate the whole predictive distribution. Thus, for variables Z and Y, we have

$$E[Z] = E[E[Z|Y]].$$

In our example, we have $E[Y|\theta] = 10\theta$ and $E[\theta|\mathbf{x}] = \frac{5}{7}$ and therefore

$$E[Y|\mathbf{x}] = 10 \times \frac{5}{7} \approx 7.141.$$

In order to calculate the predictive variance, we can use the formula

$$V[Z] = E[V[Z|Y]] + V[E[Z|Y]].$$

Beta prior distributions

As noted earlier, the uniform distribution is not a realistic prior distribution in many cases. In many cases we may have more knowledge that we can express in the form of a beta distribution.

Suppose that $\theta \sim \mathcal{B}(5,5)$. This prior might be used to express prior belief that the true probability is near to $1/2$.

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1}\theta^{9}(1-\theta)^{3} \\ &\propto \theta^{14-1}(1-\theta)^{8-1} \\ \theta|\mathbf{x} &\sim \mathcal{B}(14,8) \end{aligned}$$

Suppose now that we think that the coin is likely to be biased in favour of heads. Then we might assume that $\theta \sim \mathcal{B}(5, 1)$. In this case,

$$p(\theta|\mathbf{x}) \propto \theta^{5-1}(1-\theta)^{1-1}\theta^{9}(1-\theta)^{3}$$

and therefore, $\theta|\mathbf{x} \sim \mathcal{B}(14, 4)$.

Suppose now that we think that the coin is likely to be biased in favour of heads. Then we might assume that $\theta \sim \mathcal{B}(5, 1)$. In this case,

$$p(\theta|\mathbf{x}) \propto \theta^{5-1}(1-\theta)^{1-1}\theta^{9}(1-\theta)^{3}$$

and therefore, $\theta|\mathbf{x} \sim \mathcal{B}(14, 4)$.

We may think the coin is biased in favour of tails. In this case, we may assume that $\theta \sim \mathcal{B}(1, 5)$, when $\theta|\mathbf{x} \sim \mathcal{B}(10, 8)$.

Suppose now that we think that the coin is likely to be biased in favour of heads. Then we might assume that $\theta \sim \mathcal{B}(5, 1)$. In this case,

$$p(\theta|\mathbf{x}) \propto \theta^{5-1}(1-\theta)^{1-1}\theta^{9}(1-\theta)^{3}$$

and therefore, $\theta|\mathbf{x} \sim \mathcal{B}(14, 4)$.

We may think the coin is biased in favour of tails. In this case, we may assume that $\theta \sim \mathcal{B}(1, 5)$, when $\theta|\mathbf{x} \sim \mathcal{B}(10, 8)$.

In all 4 cases we have considered, the prior and posterior distributions are beta distributions.

Suppose now that we think that the coin is likely to be biased in favour of heads. Then we might assume that $\theta \sim \mathcal{B}(5, 1)$. In this case,

$$p(\theta|\mathbf{x}) \propto \theta^{5-1}(1-\theta)^{1-1}\theta^{9}(1-\theta)^{3}$$

and therefore, $\theta|\mathbf{x} \sim \mathcal{B}(14, 4)$.

We may think the coin is biased in favour of tails. In this case, we may assume that $\theta \sim \mathcal{B}(1, 5)$, when $\theta|\mathbf{x} \sim \mathcal{B}(10, 8)$.

In all 4 cases we have considered, the prior and posterior distributions are beta distributions.

The beta prior distribution is $conjugate$ to the binomial sampling distribution. Similar situations will be considered in chapter 3.

The scaled likelihood function

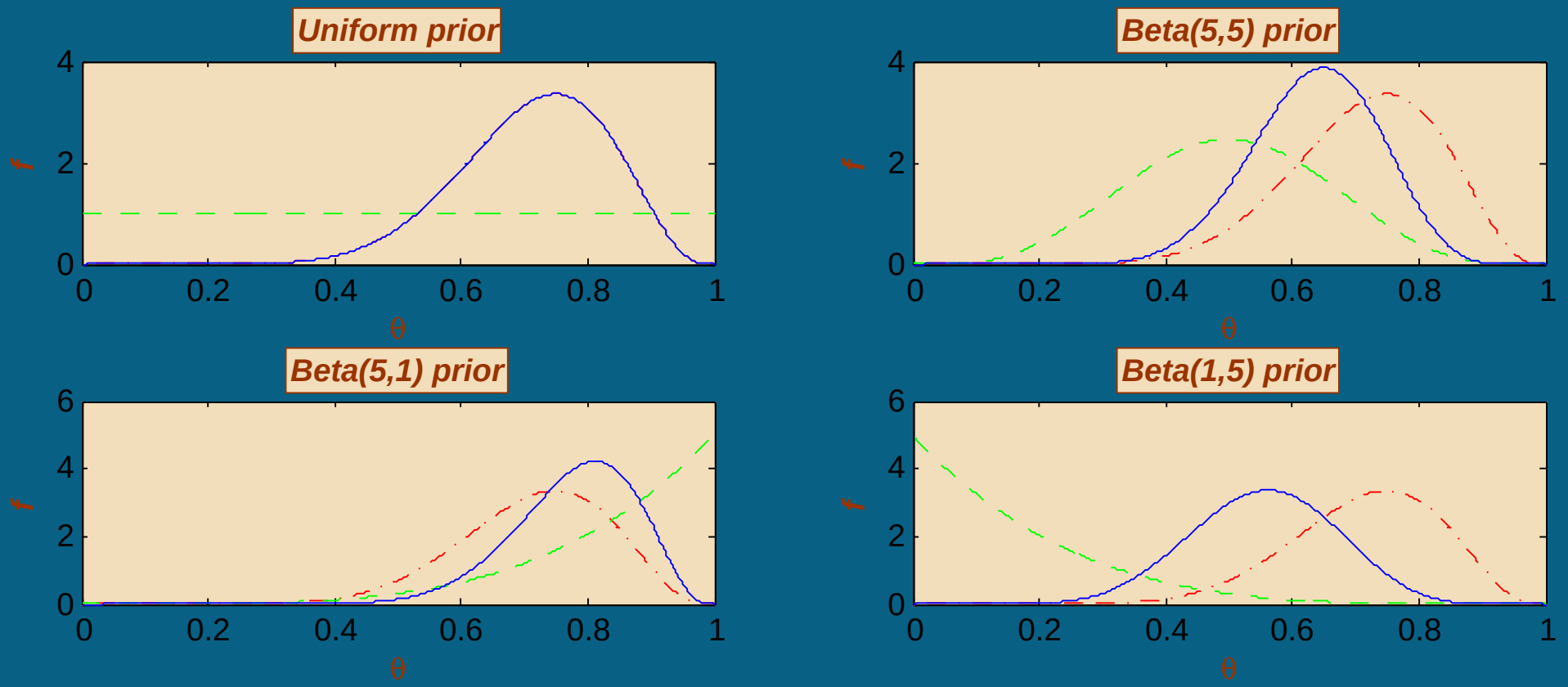
Sometimes, (if θ is one dimensional), it is possible to observe the influence of the prior distribution and the likelihood function on the posterior distribution. In order to do this, we define the scaled likelihood as

$$\frac{l(\theta|\mathbf{x})}{\int l(\theta|\mathbf{x})\ d\theta}$$

Note that the scaled likelihood does not always exist. However, when it can be defined, we can construct a diagram showing the prior, the scaled likelihood and the posterior together.

The following diagrams illustrate the effects of using the different prior distributions in this problem. In each graphic, the prior is given as a green dashed line, the scaled likelihood as a red dash-dot line and the posterior as a solid blue line.

The results of using different prior distributions



What if the prior distribution is not beta?

If we do not use a beta prior in the binomial sampling problem, then calculation of the posterior density and its properties is more complex.

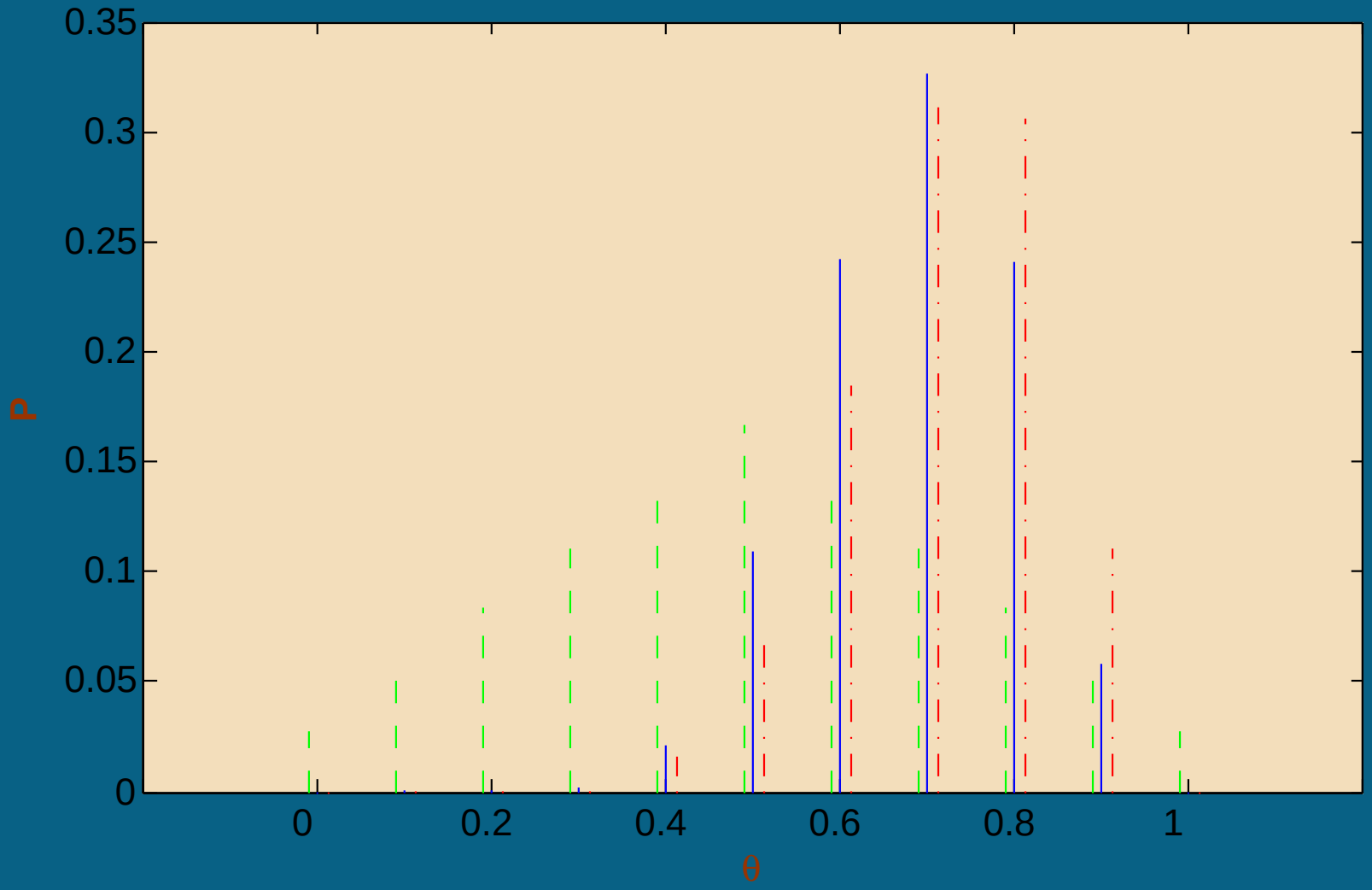
Suppose that we assume a discrete prior distribution

θ	$P(\theta)$
0	1/36
0.1	2/36
0.2	3/36
0.3	4/36
0.4	5/36
0.5	6/36
0.6	5/36
0.7	4/36
0.8	3/36
0.9	2/36
1.0	1/36

In this case, we have to use Bayes theorem to calculate all of the probabilities.

θ	$P(\theta)$	$\theta^9(1-\theta)^3$	$\frac{\theta^9(1-\theta)^3}{\sum \theta^9(1-\theta)^3}$	$P(\theta)\frac{\theta^9(1-\theta)^3}{\sum \theta^9(1-\theta)^3}$	$P(\theta\|\mathbf{x})$
0	1/36	0.0000	0.0000	0.0000	0.0000
0.1	2/36	0.0000	0.0000	0.0000	0.0000
0.2	3/36	0.0000	0.0001	0.0000	0.0001
0.3	4/36	0.0000	0.0019	0.0002	0.0020
0.4	5/36	0.0001	0.0162	0.0022	0.0212
0.5	6/36	0.0002	0.0697	0.0116	0.1097
0.6	5/36	0.0006	0.1841	0.0256	0.2415
0.7	4/36	0.0011	0.3110	0.0346	0.3263
0.8	3/36	0.0011	0.3065	0.0255	0.2412
0.9	2/36	0.0004	0.1106	0.0061	0.0580
1.0	1/36	0.0000	0.0000	0.0000	0.0000
Total	1	0.0035	1	0.1059	1.0000

The diagram shows the prior and posterior probabilities and scaled likelihood. The posterior mean can be calculated to be $E[\theta|\mathbf{x}] = 0.6824$.



Results with a continuous prior

Suppose that we use a continuous, non-beta prior, for example $\theta \sim \mathcal{N}\left(0.5, 0.2^2\right)$. In this case, the posterior distribution is

$$p(\theta|\mathbf{x}) \propto \exp\left(-\frac{(\theta-0.5)^2}{0.4}\right)\theta^9(1-\theta)^3$$

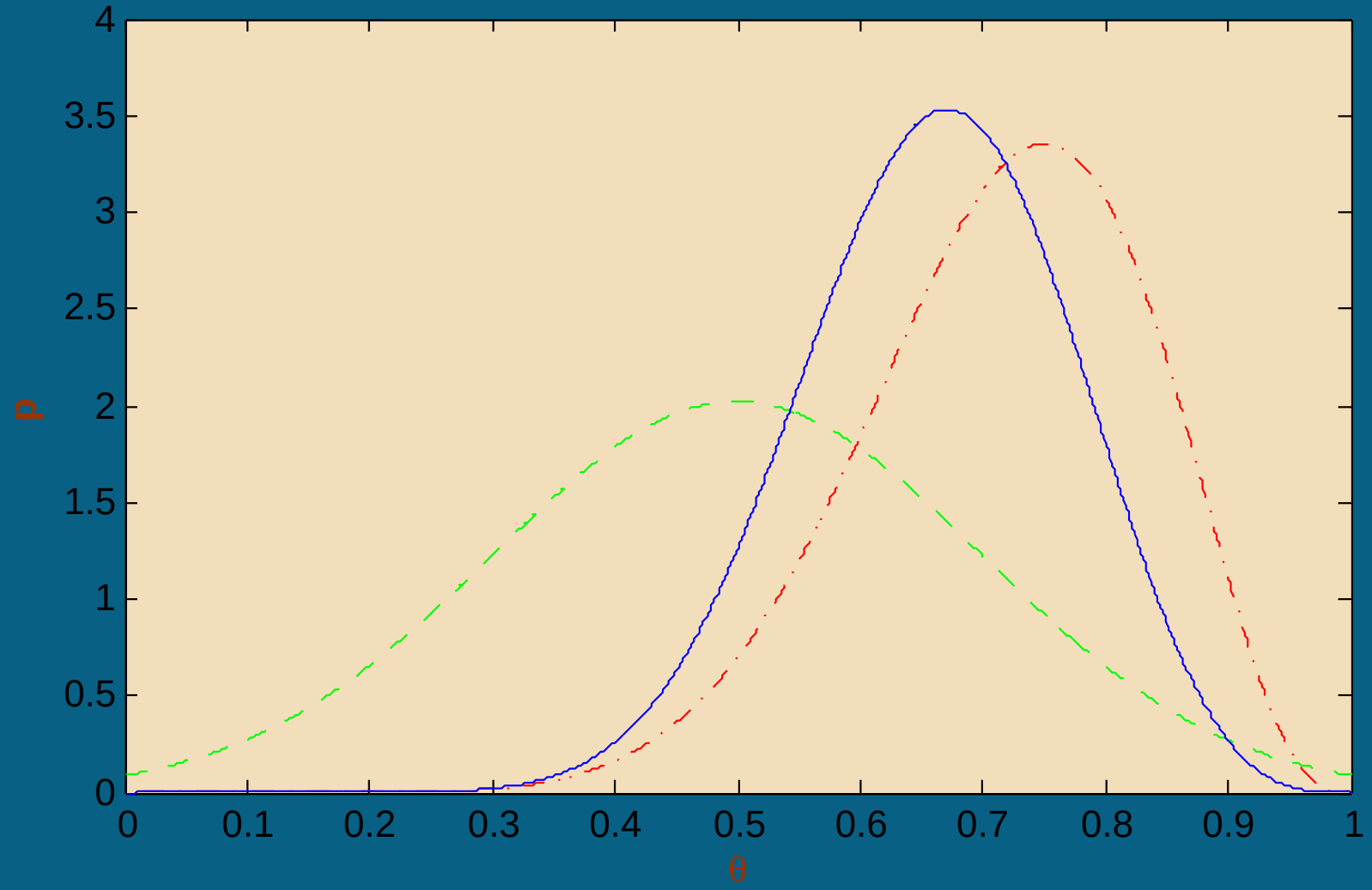
and the integration constant of this distribution must be calculated numerically.

We can do this in MATLAB by defining a function;

```
function ftheta=myfn(theta)
lik=betapdf(theta,10,4);
prior=normpdf(theta,0.5,0.2)/(normcdf(1,0.5,0.2)-normcdf(0,0.5,0.2));
ftheta=prior.*lik;
```

and using `intconst=quad(@myfn,0,1)`, which produces the result `intconst=1.1143`.

The diagram shows the prior and posterior densities and likelihood.



The posterior mean and variance etc. also have to be calculated numerically.

```
function meantheta=myfn2(theta)
lik=betapdf(theta,10,4);
prior=normpdf(theta,0.5,0.2)/(normcdf(1,0.5,0.2)-normcdf(0,0.5,0.2));
meantheta=theta.*prior.*lik;
```

Then the mean is given by `etheta=quad(@myfn2,0,1)/intconst` from which we find $E[\theta|\mathbf{x}] = 0.66$.

References

Box, G.E. and Tiao, G.C. (1992) *Bayesian Inference in Statistical Analysis* (Wiley classics library ed.). New York: Wiley.

Lindley, D.V. (1983). Theory and Practice of Bayesian Statistics. *The Statistician*, **32**, 1–11.

3. Conjugate families of distributions

Objective

One problem in the implementation of Bayesian approaches is analytical tractability. For a likelihood function $l(\boldsymbol{\theta}|\mathbf{x})$ and prior distribution $p(\boldsymbol{\theta})$, in order to calculate the posterior distribution, it is necessary to evaluate

$$f(\mathbf{x}) = \int l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

and to make predictions, we must evaluate

$$f(y|\mathbf{x}) = \int f(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

General approaches for evaluating such integrals numerically are studied in chapter 6, but in this chapter, we will study the conditions under which such integrals may be evaluated analytically.

Recommended reading

- Wikipedia entry on conjugate priors.

  http://en.wikipedia.org/wiki/Conjugate_prior

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Section 5.2.

Coin tossing problems and beta priors

It is possible to generalize what we have seen in Chapter 2 to other coin tossing situations.

Suppose that we have defined a beta prior distribution $\theta \sim \mathcal{B}(\alpha, \beta)$ for $\theta = P(\text{head})$.

Then if we observe a sample of coin toss data, whether the sampling mechanism is binomial, negative-binomial or geometric, the likelihood function always takes the form

$$l(\theta|\mathbf{x}) = c\theta^h(1-\theta)^t$$

where c is some constant that depends on the sampling distribution and h and t are the observed numbers of heads and tails respectively.

Applying Bayes theorem, the posterior distribution is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto c\theta^h(1-\theta)^t \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+h-1}(1-\theta)^{\beta+t-1} \end{aligned}$$

which implies that $\theta|\mathbf{x} \sim \mathcal{B}(\alpha+h, \beta+t)$.

Thus, using a beta prior, guarantees that the posterior distribution is also beta. In this case, we say that the class of beta prior distributions is $conjugate$ to the class of binomial (or geometric or negative binomial) likelihood functions.

Interpretation of the beta parameters

The posterior distribution is $\mathcal{B}(\alpha + h, \beta + t)$ so that α (β) in the prior plays the role of the number of heads (tails) observed in the experiment. The information represented by the prior distribution can be viewed as equivalent to the information contained in an experiment where we observe α heads and β tails.

Furthermore, the posterior mean is given by $(\alpha + h)/(\alpha + \beta + h + t)$ which can be expressed in mixture form as

$$\begin{aligned} E[\theta|\mathbf{x}] &= w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{h}{h+t} \\ &= wE[\theta] + (1-w)\hat{\theta} \end{aligned}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+h+t}$ is the relative weight of the number of equivalent coin tosses in the prior.

Limiting results

Consider also what happens as we let $\alpha, \beta \to 0$. We can interpret this as representing a prior which contains information equivalent to zero coin tosses.

In this case, the posterior distribution tends to $\mathcal{B}(h, t)$ and, for example, the posterior mean $E[\theta|\mathbf{x}] \to \hat{\theta}$, the classical MLE.

The limiting form of the prior distribution in this case is Haldane's (1931) prior,

$$p(\theta) \propto \frac{1}{\theta(1-\theta)},$$

which is an *improper* density. We analyze the use of improper densities further in chapter 5.

Prediction and posterior distribution

The form of the predictive and posterior distributions depends on the sampling distribution. We shall consider 4 cases:

- Bernoulli trials.
- Binomial data.
- Geometric data.
- Negative binomial data.

Bernoulli trials

Given that the beta distribution is conjugate in coin tossing experiments, given a (Bernoulli or binomial, etc.) sampling distribution, $f(x|\theta)$, we only need to calculate the prior predictive density $f(x) = \int f(x|\theta)p(\theta)\,d\theta$ for a beta prior. Assume first that we have a Bernoulli trial with $P(X = 1|\theta) = \theta$ and prior distribution $\theta \sim \mathcal{B}(\alpha, \beta)$. Then:

Theorem 3
If X is a Bernoulli trial with parameter θ and $\theta \sim \mathcal{B}(\alpha, \beta)$, then:

$$f(x) = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } x = 1 \\ \frac{\beta}{\alpha+\beta} & \text{if } x = 0 \end{cases}$$

$$E[X] = \frac{\alpha}{\alpha+\beta}, \qquad V[X] = \frac{\alpha\beta}{(\alpha+\beta)^2}.$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of Bernoulli data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.

Proof

$$
\begin{aligned}
P(X=1) &= \int_0^1 P(X=1|\theta)p(\theta)\,d\theta \\
&= E[\theta] = \frac{\alpha}{\alpha+\beta} \\
P(X=0) &= 1 - P(X=1) = \frac{\beta}{\alpha+\beta} \\
E[X] &= 0 \times P(X=0) + 1 \times P(X=1) = \frac{\alpha}{\alpha+\beta} \\
V[X] &= E\left[X^2\right] - E[X]^2 \\
&= \frac{\alpha}{\alpha+\beta} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2}.
\end{aligned}
$$

The posterior distribution formula was demonstrated previously. ∎

Binomial data

Theorem 4
Suppose that $X|\theta \sim \mathcal{BI}(m, \theta)$ with $\theta \sim \mathcal{B}(\alpha, \beta)$. Then, the predictive density of X is the beta binomial density

$$\begin{aligned} f(x) &= \binom{m}{x} \frac{B(\alpha + x, \beta + m - x)}{B(\alpha, \beta)} \quad \text{for } x = 0, 1, \ldots, m \\ E[X] &= m\frac{\alpha}{\alpha + \beta} \\ V[X] &= m(m + \alpha + \beta)\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

Given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, of binomial data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + \sum_{i=1}^{n} x_i, \beta + mn - \sum_{i=1}^{n} x_i)$.

Proof

$$
\begin{aligned}
f(x) &= \int_0^1 f(x|\theta)p(\theta)\,d\theta \\
&= \int_0^1 \binom{m}{x} \theta^x(1-\theta)^{m-x}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta \quad \text{for } x = 0,\ldots,m \\
&= \binom{m}{x}\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+m-x-1}\,d\theta \\
&= \binom{m}{x}\frac{B(\alpha+x,\beta+m-x)}{B(\alpha,\beta).}
\end{aligned}
$$

$$
\begin{aligned}
E[X] &= E[E[X|\theta]] \\
&= E[m\theta] \quad \text{remembering the mean of a binomial distribution} \\
&= m\frac{\alpha}{\alpha+\beta}.
\end{aligned}
$$

$$
\begin{aligned}
V[X] &= E[V[X|\theta]] + V[E[X|\theta]] \\
&= E[m\theta(1-\theta)] + V[m\theta] \\
&= m\frac{B(\alpha+1,\beta+1)}{B(\alpha,\beta)} + m^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= m\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + m^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= m(m+\alpha+\beta)\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.
\end{aligned}
$$

Geometric data

Theorem 5
Suppose that $\theta \sim \mathcal{B}(\alpha, \beta)$ and $X|\theta \sim \mathcal{GE}(\theta)$, that is

$$P(X = x|\theta) = (1-\theta)^x \theta \quad \text{for } x = 0, 1, 2, \ldots$$

Then, the predictive density of X is

$$\begin{aligned} f(x) &= \frac{B(\alpha+1, \beta+x)}{B(\alpha,\beta)} \quad \text{for } x = 0, 1, 2, \ldots \\ E[X] &= \frac{\beta}{\alpha - 1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha\beta(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, then $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

Proof Exercise. ∎

Negative binomial data

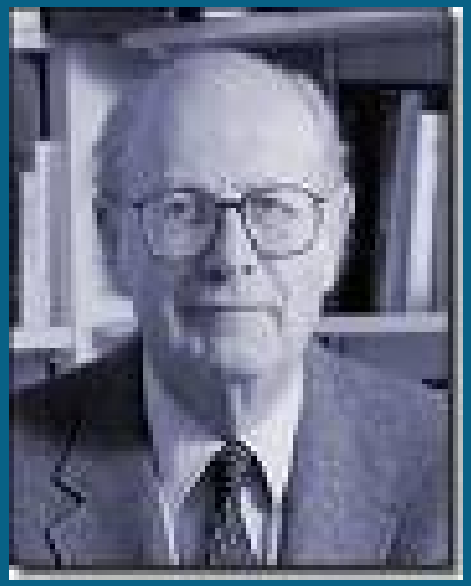
Theorem 6
Suppose that $\theta \sim \mathcal{B}(\alpha, \beta)$ and $X|\theta \sim \mathcal{NB}(r, \theta)$, that is $P(X = x|\theta) = \begin{pmatrix} x+r-1 \\ x \end{pmatrix} \theta^r (1-\theta)^x \quad \text{for } x = 0, 1, 2, \ldots$ Then:

$$\begin{aligned} f(x) &= \begin{pmatrix} x+r-1 \\ x \end{pmatrix} \frac{B(\alpha+r, \beta+x)}{B(\alpha, \beta)} \quad \text{for } x = 0, 1, \ldots \\ E[X] &= r\frac{\beta}{\alpha-1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{r\beta(r+\alpha-1)(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of negative binomial data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + rn, \beta + \sum_{i=1}^n x_i)$.

Proof Exercise. ∎

Conjugate priors



Raiffa

The concept and formal definition of conjugate prior distributions comes from Raiffa and Schlaifer (1961).

Definition 2

If $\mathcal{F}$ is a class of sampling distributions $f(x|\boldsymbol{\theta})$ and $\mathcal{P}$ is a class of prior distributions for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$ we say that $\mathcal{P}$ is **conjugate to** $\mathcal{F}$ if $p(\boldsymbol{\theta}|x) \in \mathcal{P} \; \forall \; f(\cdot|\boldsymbol{\theta}) \in \mathcal{F} \; y \; p(\cdot) \in \mathcal{P}$.

The exponential-gamma system

Suppose that $X|\theta \sim \mathcal{E}(\theta)$, is exponentially distributed and that we use a gamma prior distribution, $\theta \sim \mathcal{G}(\alpha, \beta)$, that is

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \quad \text{for} \quad \theta > 0.$$

Given sample data $\mathbf{x}$, the posterior distribution is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto p(\theta)l(\theta|\mathbf{x}) \\ &\propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\theta e^{-\theta x_i} \\ &\propto \theta^{\alpha+n-1}e^{-(\beta+n\bar{x})\theta} \end{aligned}$$

which is the nucleus of a gamma density $\theta|\mathbf{x} \sim \mathcal{G}(\alpha+n, \beta+n\bar{x})$ and thus the gamma prior is conjugate to the exponential sampling distribution.

Interpretation and limiting results

The information represented by the prior distribution can be interpreted as being equivalent to the information contained in a sample of size α and sample mean β/α.

Letting $\alpha, \beta \to 0$, then the posterior distribution approaches $\mathcal{G}(n, n\bar{x})$ and thus, for example, the posterior mean tends to $1/\bar{x}$ which is equal to the MLE in this experiment. However, the limiting prior distribution in this case is

$$f(\theta) \propto \frac{1}{\theta}$$

which is improper.

Prediction and posterior distribution

Theorem 7
Let $X|\theta \sim \mathcal{E}(\theta)$ and $\theta \sim \mathcal{G}(\alpha, \beta)$. Then the predictive density of X is

$$\begin{aligned} f(x) &= \frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}} \quad \text{for } x > 0 \\ E[X] &= \frac{\beta}{\alpha-1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of exponential data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{G}(\alpha + n, \beta + n\bar{x})$.

Proof Exercise. ∎

Exponential families

The exponential distribution is the simplest example of an exponential family distribution. Exponential family sampling distributions are highly related to the existence of conjugate prior distributions.

Definition 3
A probability density $f(x|\theta)$ where $\theta \in \mathbb{R}$ is said to belong to the one-parameter $exponential\ family$ if it has form

$$f(x|\theta) = C(\theta)h(x)\exp(\phi(\theta)s(x))$$

for given functions $C(\cdot), h(\cdot), \phi(\cdot), s(\cdot)$. If the support of X is independent of θ then the family is said to be $regular$ and otherwise it is $irregular$.

Examples of exponential family distributions

Example 12
The binomial distribution,

$$\begin{aligned} f(x|\theta) &= \binom{m}{x} \theta^x (1-\theta)^{m-x} \quad \text{for } x = 0, 1, \ldots, m \\ &= (1-\theta)^m \binom{m}{x} \exp\left(x \log \frac{\theta}{1-\theta}\right) \end{aligned}$$

is a regular exponential family distribution.

Example 13
The Poisson distribution,

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} = e^{-\theta} \frac{1}{x!} \exp(x \log \theta) \quad \text{for } x = 0, 1, 2, \ldots$$

is a regular exponential family distribution.

Irregular and non exponential family distributions

Example 14
The uniform distribution,

$$f(x|\theta) = \frac{1}{\theta} \quad \text{for } 0 < x < \theta$$

is an irregular exponential family distribution.

Example 15
The Cauchy density,

$$f(x|\theta) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} \quad \text{for} \quad -\infty < x < \infty$$

is not an exponential family distribution.

The Student's t and Fisher's F distributions and the logistic distribution are other examples of non exponential family distributions.

Sufficient statistics and exponential family distributions

Theorem 8
If $X|\theta$ is a one-parameter, regular exponential family distribution, then given a sample $\mathbf{x} = (x_1, \ldots, x_n)$, a sufficient statistic for θ is $t(\mathbf{x}) = \sum_{i=1}^{n} s(x_i)$.

Proof The likelihood function is

$$\begin{aligned} l(\theta|\mathbf{x}) &= \prod_{i=1}^{n} C(\theta)h(x_i)\exp\left(\phi(\theta)s(x_i)\right) \\ &= \left(\prod_{i=1}^{n} h(x_i)\right) C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right) \\ &= h(\mathbf{x})g(t,\theta) \end{aligned}$$

where $h(\mathbf{x}) = (\prod_{i=1}^{n} h(x_i))$ and $g(t,\theta) = C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right)$ and therefore, t is a sufficient statistic as in Theorem 1. ∎

A conjugate prior to an exponential family distribution

If $f(x|\theta)$ is an exponential family, with density as in Definition 3, then a conjugate prior distribution for θ exists.

Theorem 9
The prior distribution $p(\theta) \propto C(\theta)^a \exp\left(\phi(\boldsymbol{\theta})b\right)$ is conjugate to the exponential family distribution likelihood.

Proof From Theorem 8, the likelihood function for a sample of size n is

$$l(\theta|\mathbf{x}) \propto C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right)$$

and given $p(\boldsymbol{\theta})$ as above, we have a posterior of the same form

$$p(\theta|\mathbf{x}) \propto C(\theta)^{a^\star} \exp\left(\phi(\theta)b^\star\right)$$

where $a^\star = a + n$ and $b^\star = b + t(\mathbf{x})$ for $j = 1, \ldots, n$. ∎

The Poisson-gamma system

Suppose that $X|\theta \sim \mathcal{P}(\theta)$. Then, from Example 13, we have the exponential family form

$$f(x|\theta) = e^{-\theta}\frac{1}{x!}\exp(x\log\theta)$$

and from Theorem 9, a conjugate prior density for θ has form

$$p(\theta) \propto \left(e^{-\theta}\right)^{a}\exp(b\log\theta)$$

for some a, b. Thus, $p(\theta) \propto \theta^{b}e^{-a\theta}$ which is the nucleus of a gamma density, $\theta \sim \mathcal{G}(\alpha, \beta)$, where $\alpha = b + 1$ and $\beta = a$.

Thus, the gamma prior distribution is conjugate to the Poisson sampling distribution.

Derivation of the posterior distribution

Now assume the prior distribution $\theta \sim \mathcal{G}(\alpha, \beta)$ and suppose that we observe a sample of n Poisson data. Then, we can derive the posterior distribution via Bayes theorem as earlier.

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!} \\ &\propto \theta^{\alpha+n\bar{x}-1}e^{-(\beta+n)\theta} \\ \theta|\mathbf{x} &\sim \mathcal{G}(\alpha+n\bar{x}, \beta+n) \end{aligned}$$

Derivation of the posterior distribution

Now assume the prior distribution $\theta \sim \mathcal{G}(\alpha, \beta)$ and suppose that we observe a sample of n Poisson data. Then, we can derive the posterior distribution via Bayes theorem as earlier.

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!} \\ &\propto \theta^{\alpha+n\bar{x}-1}e^{-(\beta+n)\theta} \\ \theta|\mathbf{x} &\sim \mathcal{G}(\alpha+n\bar{x}, \beta+n) \end{aligned}$$

Alternatively, we can use Theorem 9. The sufficient statistic is $t(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and thus, from Theorem 9, we know immediately that

$$\theta|\mathbf{x} \sim \mathcal{G}(\underbrace{\alpha}_{b+1}+t(\mathbf{x}), \underbrace{\beta}_{a}+n) \sim \mathcal{G}(\alpha+n\bar{x}, \beta+n).$$

Interpretation and limiting results

We might interpret the information in the prior as being equivalent to a the information contained in a sample of size β with sample mean α/β.

Letting $\alpha, \beta \to 0$, the posterior mean converges to the MLE although the corresponding limiting prior, $p(\theta) \propto \frac{1}{\theta}$, is improper.

Prediction and posterior distribution

Theorem 10
Let $X|\theta \sim \mathcal{P}(\theta)$ with $\theta \sim \mathcal{G}(\alpha, \beta)$. Then:

$$\begin{aligned} X &\sim \mathcal{NB}\left(\alpha, \frac{\beta}{\beta+1}\right) \quad \text{that is,} \\ f(x) &= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x} \quad \text{for } x = 0, 1, 2, \ldots \\ E[X] &= \frac{\alpha}{\beta} \\ V[X] &= \frac{\alpha(\beta+1)}{\beta^2}. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of Poisson data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{G}(\alpha + n\bar{x}, \beta + n)$.

Proof Exercise. ∎

The uniform-Pareto system

As noted in Example 14, the uniform distribution is not a regular exponential family distribution. However, we can still define a conjugate prior distribution.

Let $X \sim \mathcal{U}(0, \theta)$. Then, given a sample of size n, the likelihood function is $l(\theta|\mathbf{x}) = \frac{1}{\theta^n}, \quad \text{for } \theta > x_{\max}$, where $x_{\max}$ is the sample maximum.

Consider a $Pareto$ prior distribution, $\theta \sim \mathcal{PA}(\alpha, \beta)$, with density

$$p(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \quad \text{for } \theta > \beta > 0$$

and mean $E[\theta] = \alpha\beta/(\alpha - 1)$.

It is clear that this prior distribution is conjugate and that

$$\theta|\mathbf{x} \sim \mathcal{PA}\left(\alpha^\star, \beta^\star\right)$$

where $\alpha^\star = \alpha + n$ and $\beta^\star = \max\{\beta, x_{\max}\}$.

Properties and limiting results

The posterior mean in this case is

$$E[\theta|\mathbf{x}] = \frac{(\alpha+n)\max\{\beta, x_{\max}\}}{\alpha+n-1}$$

which cannot be represented in a general way as an average of the prior mean and the MLE ($x_{\max}$).

We can interpret the information in the prior as being equivalent to a sample of size α where the maximum value is equal to β. When we let $\alpha, \beta \longrightarrow 0$, then the posterior distribution approaches $\mathcal{PA}(n, x_{max})$ and in this case, the posterior mean is

$$E[\theta|\mathbf{x}] = \frac{nx_{\max}}{n-1} > x_{\max} = \hat{\theta}.$$

This also corresponds to an improper limiting prior $p(\theta) \propto 1/\theta$.

In this experiment, no prior distribution (except for a point mass at $x_{\max}$) will lead to a posterior distribution whose mean coincides with the MLE.

Prediction and posterior distribution

Theorem 11
If $X|\theta \sim \mathcal{U}(0,\theta)$ and $\theta \sim \mathcal{PA}(\alpha,\beta)$, then:

$$\begin{aligned} f(x) &= \begin{cases} \frac{\alpha}{(\alpha+1)\beta} & \text{if } 0 < x < \beta \\ \frac{\alpha\beta^{\alpha}}{(\alpha+1)x^{\alpha+1}} & \text{if } x \geq \beta \end{cases} \\ E[X] &= \frac{\alpha\beta}{2(\alpha-1)} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha(\alpha+2)\beta^2}{12(\alpha-1)(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample of size n, then the posterior distribution of θ is $\theta|\mathbf{x} \sim \mathcal{PA}\left(\alpha+n, \max\{\beta, x_{\max}\}\right)$.

Proof Exercise. ∎

Higher dimensional exponential family distributions

Definition 4

A probability density $f(\mathbf{x}|\boldsymbol{\theta})$ where $\theta \in \mathbb{R}^k$ is said to belong to the k-parameter *exponential family* if it has form

$$f(\mathbf{x}|\boldsymbol{\theta}) = C(\boldsymbol{\theta})h(\mathbf{x}) \exp\left(\sum_{j=1}^{k} \phi_j(\boldsymbol{\theta})s_j(\mathbf{x})\right)$$

for given functions $C(\cdot), h(\cdot), \phi(\cdot), s(\cdot)$. If the support of $\mathbf{X}$ is independent of θ then the family is said to be *regular* and otherwise it is *irregular*.

Example 16

The multinomial density is given by $f(\mathbf{x}|\boldsymbol{\theta}) = \frac{m!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k} \theta_j^{x_j}$ where $x_j \in \{0, 1, \ldots, m\}$ for $j = 1, \ldots, k$, $\sum_{j=1}^{k} x_j = m$ and $\sum_{j=1}^{k} \theta_j = 1$.

We can write

$$
\begin{aligned}
f(\mathbf{x}|\boldsymbol{\theta}) &= \frac{m!}{\prod_{j=1}^{k} x_j!} \exp\left(\sum_{j=1}^{k} x_j \log(\theta_j)\right) \\
&= \frac{m!}{\prod_{j=1}^{k} x_j!} \exp\left((m - \sum_{j=1}^{k-1} x_j) \log \theta_k + \sum_{j=1}^{k-1} x_j \log(\theta_j)\right) \\
&= \frac{m!}{\prod_{j=1}^{k} x_j!} \theta_k^m \exp\left(\sum_{j=1}^{k-1} x_j \log(\theta_j/\theta_k)\right)
\end{aligned}
$$

and thus, the multinomial distribution is a regular, $k-1$ dimensional exponential family distribution.

It is clear that we can generalize Theorem 8 to the k dimensional case.

Theorem 12
If $\mathbf{X}|\boldsymbol{\theta}$ is a k-parameter, regular exponential family distribution, then given a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, a sufficient statistic for $\boldsymbol{\theta}$ is $\mathbf{t}(\mathbf{x}) = (\sum_{i=1}^n s_1(\mathbf{x}_i), \ldots, \sum_{i=1}^n s_k(\mathbf{x}_i))$.

Proof Using the same arguments as in Theorem 8, the result is easy to derive.

A conjugate prior to an exponential family sampling distribution

We can also generalize Theorem 9. If $f(x|\boldsymbol{\theta})$ is an exponential family, with density as in Definition 4, then a conjugate prior distribution for $\boldsymbol{\theta}$ exists.

Theorem 13
The prior distribution $p(\boldsymbol{\theta}) \propto C(\boldsymbol{\theta})^a \exp\left(\sum_{j=1}^{k} \phi_j(\boldsymbol{\theta}) b_j\right)$ is conjugate to the k dimensional exponential family distribution likelihood.

Proof Exercise. ∎

The multinomial-Dirichlet system

Suppose that $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ is a k- dimensional multinomial sampling distribution. Then from Example 16, we can express the multinomial density as

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{m!}{\prod_{j=1}^{k} x_j!} \theta_k^m \exp\left(\sum_{j=1}^{k-1} x_j \log(\theta_j/\theta_k)\right)$$

and therefore, a conjugate prior distribution is

$$\begin{aligned} f(\boldsymbol{\theta}) &\propto (\theta_k{}^m)^a \exp\left(\sum_{j=1}^{k-1} b_j \log(\theta_j/\theta_k)\right) \quad \text{for} \quad \text{arbitrary } a, b_1, \ldots, b_{k-1} \\ &\propto \prod_{j=1}^{k-1} \theta_j{}^{b_j} \theta_k^{a-\sum_{j=1}^{k-1} b_j} \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \end{aligned}$$

where $\alpha_j = b_j + 1$ for $j = 1, \ldots, k-1$ and $\alpha_k = a - \sum_{j=1}^{k} b_j + 1$.

The Dirichlet distribution

Definition 5

A random variable, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is said to have a Dirichlet distribution, $\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ if

$$p(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

for $0 < \theta_j < 1$, $\sum_{j=1}^{k} \theta_j = 1$.

The Dirichlet distribution may be thought of as a generalization of the beta distribution. In particular, the marginal distribution of any θ_j is beta, that is $\theta_j \sim \mathcal{B}(\alpha_j, \alpha_0 - \alpha_j)$, where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.

Also, the moments of the Dirichlet distribution are easily evaluated:

$$
\begin{aligned}
E[\theta_j] &= \frac{\alpha_j}{\alpha_0} \\
V[\theta_j] &= \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \\
Cov[\theta_i \theta_j] &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}
\end{aligned}
$$

Prediction and posterior distribution

Theorem 14
Let $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$. Then:

$$P(\mathbf{X} = \mathbf{x}) = \frac{m!\Gamma(\alpha_0)}{\Gamma(m + \alpha_0)} \prod_{j=1}^{k} \frac{\Gamma(x_j + \alpha_j)}{x_j!\Gamma(\alpha_j)} \quad \text{for } x_j \geq 0, \sum_{j=1}^{k} x_j = m$$

where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$. Also

$$\begin{aligned} E[X_j] &= m\frac{\alpha_j}{\alpha_0} \\ V[X_j] &= m(\alpha_0 + m)\frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \\ Cov[X_i, X_j] &= -m(\alpha_0 + m)\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

Given a sample, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of multinomial data, then

$$\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{D}\left(\alpha_1 + \sum_{j=1}^{n} x_{1j}, \ldots, \alpha_k + \sum_{j=1}^{n} x_{kj}\right).$$

Proof Exercise ∎

Canonical form exponential family distributions

Suppose that we have a standard exponential family distribution

$$f(\mathbf{x}|\boldsymbol{\theta}) = C(\boldsymbol{\theta})h(\mathbf{x})\exp\left(\sum_{j=1}^{k}\phi_j(\boldsymbol{\theta})s_j(\mathbf{x})\right).$$

Define the transformed variables $Y_j = s_j(\mathbf{X})$ and transformed parameters $\phi_j = \phi_j(\boldsymbol{\theta})$ for $j = 1, \ldots k$. Then the density of $\mathbf{y}|\boldsymbol{\phi}$ is of *canonical form*.

Definition 6

The density

$$f(\mathbf{y}|\boldsymbol{\phi}) = D(\mathbf{y})\exp\left(\sum_{j=1}^{k} y_j\phi_j - e(\boldsymbol{\phi})\right)$$

is called the canonical form representation of the exponential family distribution.

Conjugate analysis

Clearly, a conjugate prior for $\boldsymbol{\phi}$ takes the form

$$p(\boldsymbol{\phi}) \propto \exp\left(\sum_{j=1}^{k} b_j \phi_j - a e(\boldsymbol{\phi})\right) \propto \exp\left(\sum_{j=1}^{k} a B_j \phi_j - a e(\boldsymbol{\phi})\right)$$

where $B_j = \frac{b_j}{a}$ for $j = 1, \ldots, k$. We shall call this form the *canonical prior* for $\boldsymbol{\phi}$. If a sample of size n is observed, then

$$\begin{aligned} p(\boldsymbol{\phi}|\mathbf{y}) &\propto \exp\left(\sum_{j=1}^{k} \frac{a B_j + n\bar{y}_{\cdot j}}{a+n} \phi_j - (a+n) e(\boldsymbol{\phi})\right) \\ &\propto \exp\left(\frac{a\mathbf{B} + n\bar{\mathbf{y}}^T}{a+n}\, \boldsymbol{\phi} - (a+n) e(\boldsymbol{\phi})\right) \end{aligned}$$

The updating process for conjugate models involves a simple weighted average representation.

The posterior mean as a weighted average

Suppose that we observe a sample of n data from a canonical exponential family distribution as in Definition 6 and that we use a canonical prior distribution

$$p(\boldsymbol{\phi}) \propto \exp\left(\sum_{j=1}^{k} aB_j\phi_j - ae(\boldsymbol{\phi})\right) \propto \exp\left(a\mathbf{B}^T\boldsymbol{\phi} - ae(\boldsymbol{\phi})\right).$$

Then we can demonstrate the following theorem.

Theorem 15
$E\left[\nabla e(\boldsymbol{\phi})\right] = \frac{a\mathbf{B}+n\bar{\mathbf{y}}}{a+n}$ where $\left[\nabla e(\boldsymbol{\phi})\right]_j = \frac{\partial}{\partial\phi_j}e(\boldsymbol{\phi})$.

Proof As we have shown that the prior is conjugate, it is enough to prove that, a priori, $E[\nabla e(\boldsymbol{\phi})] = \mathbf{B}$. However,

$$a\left(\mathbf{B} - E[\nabla e(\boldsymbol{\phi})]\right) = \int a\left(\mathbf{B} - \nabla e(\boldsymbol{\phi})\right)p(\boldsymbol{\phi})\,d\boldsymbol{\phi} = \int \nabla p(\boldsymbol{\phi})\,d\boldsymbol{\phi}$$

∎

Example 17
Let $X|\theta \sim \mathcal{P}(\theta)$. Then, from Example 13, we can write

$$f(x|\theta) = e^{-\theta}\frac{1}{x!}\exp(x\log\theta) = \frac{1}{x!}\exp(x\log\theta - \theta) = \frac{1}{x!}\exp\left(x\phi - e^{\phi}\right),$$

where $\phi = e^{\theta}$, in canonical exponential family form. We already know that the gamma prior density $\theta \sim \mathcal{G}(\alpha, \beta)$ is conjugate here. Thus,

$$\begin{aligned} p(\theta) &\propto \theta^{\alpha-1}e^{-\beta\theta} \Rightarrow \\ p(\phi) &\propto (\log\phi)^{\alpha}\phi^{-\beta} \\ &\propto \exp\left(\beta\frac{\alpha}{\beta}\phi - \beta e^{\phi}\right) \end{aligned}$$

in canonical form. Now $\nabla\phi = \frac{d}{d\phi}e^{\phi} = e^{\phi} = \theta$. Thus, from Theorem 15, we have $E[\theta|\mathbf{x}] = \frac{\beta}{\beta+n}\frac{\alpha}{\beta} + \frac{n}{\alpha+n}\bar{x}$, a weighted average of the prior mean and the MLE.

Example 18

Consider the Bernoulli trial $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$. We have

$$\begin{aligned} f(x|\theta) &= (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \\ &= \exp\left(x \log\frac{\theta}{1-\theta} - \log\frac{1}{1-\theta}\right) \\ &= \exp\left(x\phi - \log(1+e^\phi)\right) \end{aligned}$$

in canonical form, where $\phi = \log\frac{\theta}{1-\theta}$.

Thus, the canonical prior for ϕ must take the form ,

$$\begin{aligned} p(\phi) &\propto \exp\left(aB\phi - a\log(1+e^\phi)\right) \\ &\propto (1+e^\phi)^{-a}\exp(aB\phi). \end{aligned}$$

This implies, using the standard change of variables formula, that

$$\begin{aligned} p(\theta) &\propto (1-\theta)^a \left(\frac{\theta}{1-\theta}\right)^{aB} \frac{1}{\theta(1-\theta)} \\ &\propto \theta^{aB-1}(1-\theta)^{a(1-B)-1} \end{aligned}$$

which we can recognize as a beta distribution, $\mathcal{B}(\alpha, \beta)$, with parameters $\alpha = aB$ and $\beta = a(1-B)$.

Now $\nabla \log(1+e^\phi) = \frac{e^\phi}{1+e^\phi} = \theta$ and so, from Theorem 15, given a sample of n Bernoulli trials, we have

$$E[\theta|\mathbf{x}] = \frac{aB + n\bar{x}}{a+n} = w\frac{\alpha}{\alpha+\beta} + (1-w)\bar{x}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+n}$.

We previously derived this formula by standard Bayesian analysis on page 95.

Mixtures of conjugate priors

Suppose that a simple conjugate prior distribution $p(\cdot) \in \mathcal{P}$ does not well represent our prior beliefs. An alternative is to consider a mixture of conjugate prior distributions

$$\sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta})$$

where $0 < w_i < 1$ and $\sum_{i=1}^{k} w_i = 1$ and $p_i(\cdot) \in \mathcal{P}$.

Note that Dalal and Hall (1983) demonstrate that any prior density for an exponential family can be approximated arbitrarily closely by a mixture of conjugate distributions.

It is easy to demonstrate that given a conjugate prior mixture distribution the posterior distribution is also a mixture of conjugate densities.

Proof Using Bayes theorem, we have

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{x}) &\propto f(\mathbf{x}|\boldsymbol{\theta}) \sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta}) \\
&\propto \sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) \\
&\propto \sum_{i=1}^{k} w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \frac{p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta})}{\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \\
&\propto \sum_{i=1}^{k} w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right) p_i(\boldsymbol{\theta}|\mathbf{x})
\end{aligned}$$

where $p_i(\cdot|\mathbf{x}) \in \mathcal{P}$ because it is assumed that $p_i(\cdot)$ is conjugate and therefore $p(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{k} w_i^\star p_i(\boldsymbol{\theta}|\mathbf{x})$ where $w_i^\star = \frac{w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right)}{\sum_{j=1}^{k} w_j \left(\int p_j(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right)}$. ∎

Example 19

In Example 11, suppose that we use a mixture prior distribution:

$$\theta \sim 0.25\mathcal{B}(1,1) + 0.75\mathcal{B}(5,5).$$

Then, the posterior distribution is given by

$$\begin{aligned}
p(\theta|\mathbf{x}) &\propto \left(0.25 \times 1 + 0.75 \times \frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1}\right)\theta^9(1-\theta)^3 \\
&\propto 0.25\theta^{10-1}(1-\theta)^{4-1} + 0.75\frac{1}{B(5,5)}\theta^{14-1}(1-\theta)^{8-1} \\
&\propto B(10,4)\frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} + 3\frac{B(14,8)}{B(5,5)}\frac{1}{B(14,8)}\theta^{14-1}(1-\theta)^{8-1} \\
&= w^\star\frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} + (1-w^\star)\frac{1}{B(14,8)}\theta^{14-1}(1-\theta)^{8-1}
\end{aligned}$$

where $w^\star = \frac{B(10,4)}{B(10,4)+3B(14,8)/B(5,5)} = 0.2315$.

Thus, the posterior distribution is a mixture of $\mathcal{B}(10,4)$ and $\mathcal{B}(14,8)$ densities with weights 0.2315 and 0.7685 respectively.

Software for conjugate models

Some general software for undertaking conjugate Bayesian analysis has been developed.

- First Bayes is a slightly dated but fairly complete package
- The book on Bayesian computation by Albert (2009) gives a number of R routines for fitting conjugate and non conjugate models all contained in the R LearnBayes package.

References

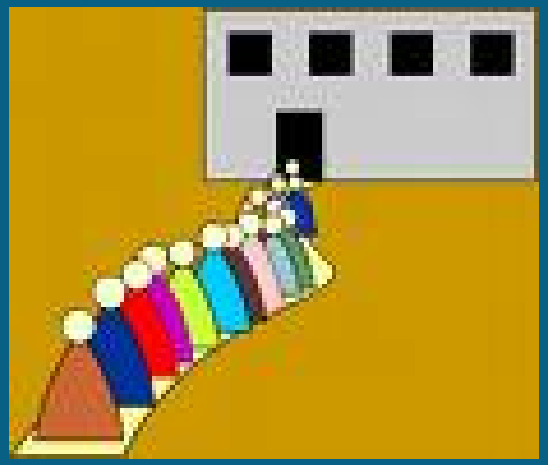
Albert, J. (2009) Bayesian Computation with R. Berlin: Springer.

Dalal, S.R. and Hall, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society Series B*, **45**, 278–286.

Haldane, J.B.S. (1931). A note on inverse probability. *Proceedings of the Cambridge Philosophical Society*, **28**, 55–61.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Cambridge MA: Harvard University Press.

Application I: Bayesian inference for Markovian queuing systems



A queue of children

We will study inference for the $M/M/1$ queueing system, as developed by Armero and Bayarri (1994a).

The $M/M/1$ queueing system

The $M(\lambda)/M(\mu)/1$ system is a queue with a single server and exponential inter-arrival and service times with rates λ and μ respectively.

Under this model, clients arrive at the checkout according to a Markov process with rate λ and if the server is empty, they start an exponential service time process with rate μ. If the server is busy, the client joins the queue of customers waiting to be served.

Stability and equilibrium distributions

In practical analysis, it is of particular importance to study the conditions under which such a queueing system is stable, i.e. when the queue length does not go off to infinity as more customers arrive. It can be shown that the system is stable if the *traffic intensity*, defined by $\rho = \lambda/\mu$ is strictly less than 1.

In the case that the system is stable, it is interesting to consider the equilibrium or stationary distribution of the queue size, client waiting times etc.

It can be shown that the equilibrium distribution of the number of clients in the system, N is geometric;

$$P(N = n|\rho) = (1-\rho)\rho^n \quad \text{for } N = 0, 1, 2, \ldots \quad \text{with mean} E[N|\rho] = \frac{\rho}{1-\rho}.$$

Also, the limiting distribution of the time, W, spent in the system by a customer is exponential; $W|\lambda, \mu \sim \mathcal{E}(\mu - \lambda)$, with mean $E[W|\lambda, \mu] = \frac{1}{\mu-\lambda}$.

The distributions of other variables of interest such as the duration of a busy period are also well known. See e.g. Gross and Harris (1985).

Experiment and inference

In real systems, the arrival and service rates will be unknown and we must use inferential techniques to estimate these parameters and the system characteristics. A first step is to find a reasonable way of collecting data.

The simplest experiment is that of observing n_l inter-arrival times and n_s service times. In this case, the likelihood is

$$l(\lambda, \mu | \mathbf{x}, \mathbf{y}) \propto \lambda^{n_l} e^{-\lambda t_l} \mu^{n_s} e^{-\mu t_s}$$

where t_l and t_s are the sums of inter-arrival and service times respectively.

If we use conjugate, independent gamma prior distributions,

$$\lambda \sim \mathcal{G}(\alpha_l, \beta_l) \quad \mu \sim \mathcal{G}(\alpha_s, \beta_s),$$

then the posterior distributions are also gamma distributions:

$$\lambda | \mathbf{x} \sim \mathcal{G}(\alpha_l + n_l, \beta_l + t_l) \quad \mu | \mathbf{y} \sim \mathcal{G}(\alpha_s + n_s, \beta_s + t_s).$$

Estimation of the traffic intensity

It is straightforward to estimate the mean of the traffic intensity ρ. We have

$$\begin{aligned} E[\rho|\mathbf{x},\mathbf{y}] &= E\left[\frac{\lambda}{\mu}\middle|\mathbf{x},\mathbf{y}\right] \\ &= E[\lambda|\mathbf{x}]E\left[\frac{1}{\mu}\middle|\mathbf{y}\right] \\ &= \frac{\alpha_l+n_l}{\beta_l+t_l}\frac{\beta_s+t_s}{\alpha_s+n_s-1} \end{aligned}$$

We can also evaluate the distribution of ρ by recalling that if $\phi \sim \mathcal{G}\left(\frac{a}{2},\frac{b}{2}\right)$, then $b\phi \sim \chi^2_a$ is chi-square distributed.

The probability that the system is stable

Thus, we have

$$2(\beta_l + t_l)\lambda|\mathbf{x} \sim \chi^2_{2(\alpha_l+n_l)} \quad 2(\beta_s + t_s)\mu|\mathbf{y} \sim \chi^2_{2(\alpha_s+n_s)}$$

and therefore, recalling that the ratio of two independent χ^2 variables divided by their degrees of freedom is an F distributed variable, we have

$$\frac{(\beta_l + t_l)(\alpha_s + n_s)}{(\alpha_l + n_l)(\beta_s + t_s)}\rho|\mathbf{x}, \mathbf{y} \sim \mathcal{F}^{2(\alpha_l+n_l)}_{2(\alpha_s+n_s)}.$$

The posterior probability that the system is stable is

$$\begin{aligned} p &= P(\rho < 1|\mathbf{x}, \mathbf{y}) \\ &= P\left(F < \frac{(\beta_l + t_l)(\alpha_s + n_s)}{(\alpha_l + n_l)(\beta_s + t_s)}\right) \quad \text{where } F \sim \mathcal{F}^{2(\alpha_l+n_l)}_{2(\alpha_s+n_s)} \end{aligned}$$

which can be easily evaluated in e.g. `Matlab` or `R`.

Estimation of the queue size in equilibrium

If p is large, it is natural to suppose that the system is stable. In this case, it is interesting to carry out prediction for the equilibrium distribution of the number of clients in the system. We have

$$\begin{aligned} P(N = n|\mathbf{x}, \mathbf{y}, \text{equilibrium}) &= P(N = n|\mathbf{x}, \mathbf{y}, \rho < 1) \\ &= \int_0^1 (1-\rho)\rho^n p(\rho|\mathbf{x}, \mathbf{y}, \rho < 1)\, d\rho \\ &= \frac{1}{p}\int_0^1 (1-\rho)\rho^n p(\rho|\mathbf{x}, \mathbf{y})\, d\rho \end{aligned}$$

We will typically need numerical methods to evaluate this integral. One possibility is to use simple *numerical integration*. Another technique is to use *Monte Carlo sampling*.

Aside: Monte Carlo sampling

Suppose that we wish to estimate an integral

$$E[g(X)] = \int g(x) f(x)\, dx$$

where X is a random variable with density $f(\cdot)$. Then, if we draw a sample, say $x_1, \ldots, x_M$ of size from $f(\cdot)$ then, under certain regularity conditions, as $M \to \infty$, we have

$$\bar{g} = \frac{1}{M} \sum_{i=1}^{M} g(x_i) \to E[g(X)].$$

In order to assess the precision of a Monte Carlo based estimator, a simple technique is to calculate the confidence band $\bar{g} \pm 2s_g/\sqrt{M}$ where $s_g^2 = \frac{1}{M}\sum_{i=1}^{M} \left(g(x_i) - \bar{g}\right)^2$ is the sample variance of $g(\cdot)$. The Monte Carlo sample size can be increased until the size of the confidence band goes below a certain preset precision ϵ.

We will discuss Monte Carlo methods in more detail in chapter 6.

Using Monte Carlo to estimate the system size distribution

We can set up a generic algorithm to generate a Monte Carlo sample.

1. Fix a large value M.

2. For $i = 1, \ldots, M$:

   (a) Generate $\lambda_i \sim \mathcal{G}(\alpha_l + n_l, \beta_l + n_l)$ and $\mu_i \sim \mathcal{G}(\alpha_s + n_s, \beta_s + n_s)$.
   (b) Set $\rho_i = \lambda_i / \mu_i$.
   (c) If $\rho_i > 1$, go to a).

Given the Monte Carlo sampled data, we can then estimate the queue size probabilities, $P(N = n | \mathbf{x}, \mathbf{y}, \text{equilibrium}) \approx \frac{1}{M} \sum_{i=1}^{M} (1 - \rho_i) \rho_i^n$ for $n = 0, 1, 2, \ldots$ and the waiting time distribution, $P(W \leq w | \mathbf{x}, \mathbf{y}, \text{equilibrium}) \approx 1 - \frac{1}{M} \sum_{i=1}^{M} e^{-(\mu_i - \lambda_i)}$.

Estimating the mean of the equilibrium system size distribution

It is easier to evaluate the predictive moments of N. We have

$$\begin{aligned} E[N|\mathbf{x},\mathbf{y},\text{equilibrium}] &= E[E[N|\rho]|\mathbf{x},\mathbf{y},\rho<1] \\ &= E\left[\frac{\rho}{1-\rho}\middle|\mathbf{x},\mathbf{y},\rho<1\right] \\ &= \frac{1}{p}\int_0^1 \frac{\rho}{1-\rho}p(\rho|\mathbf{x},\mathbf{y})\,d\rho = \infty \end{aligned}$$

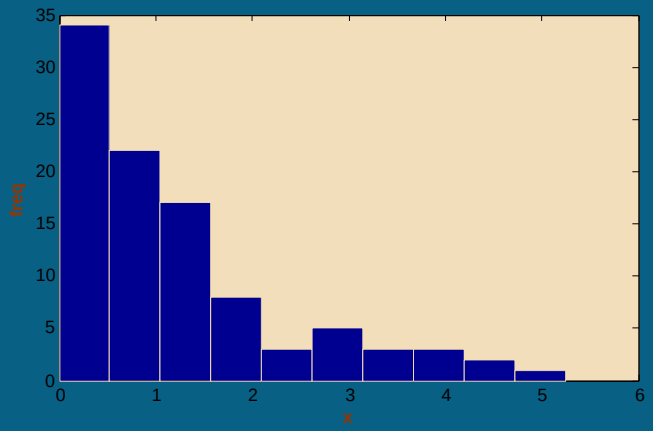
and thus, these moments do not exist.

This is a general characteristic of inference in queueing systems with Markovian inter-arrival or service processes. The same thing also happens with the equilibrium waiting time or busy period distributions. See Armero and Bayarri (1994a) or Wiper (1998).

The problem stems from the use of (independent) prior distributions for λ and μ with $p(\lambda=\mu)>0$ and can be partially resolved by assuming *a priori* that $P(\lambda\geq\mu)=0$. See Armero and Bayarri (1994b) or Ruggeri et al (1996).

Example

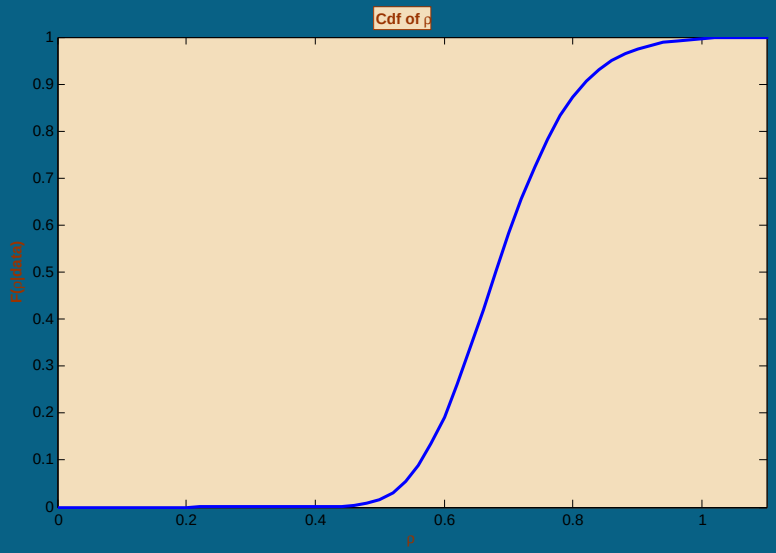
Hall (1991) gives collected inter-arrival and service time data for 98 users of an automatic teller machine in Berkeley, USA. We shall assume here that inter-arrival and service times can be modeled as exponential variables.



The sufficient statistics are $n_a = n_s = 98$, $t_a = 119.71$ and $t_s = 81.35$ minutes respectively. We will assume the $improper$ prior distributions $p(\lambda) \propto \frac{1}{\lambda}$ and $p(\mu) \propto \frac{1}{\mu}$ which we have seen earlier as limiting cases of the conjugate gamma priors. Then $\lambda|\mathbf{x} \sim \mathcal{G}(98, 119.71)$ and $\mu|\mathbf{y} \sim \mathcal{G}(98, 119.71)$.

The posterior distribution of the traffic intensity

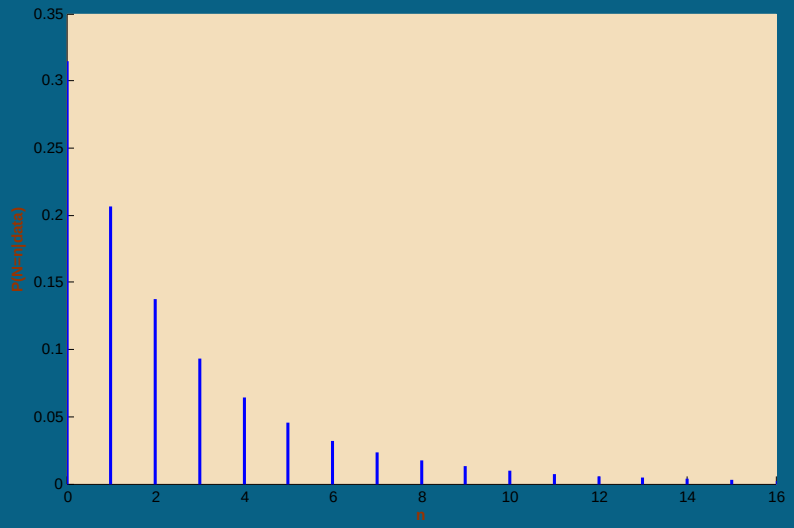
The posterior expected value of ρ is $E[\rho|\mathbf{x},\mathbf{y}] \approx 0.69$ and the distribution function $F(\rho|\mathbf{x},\mathbf{y})$ is illustrated below.



The posterior probability that the system is stable is 0.997.

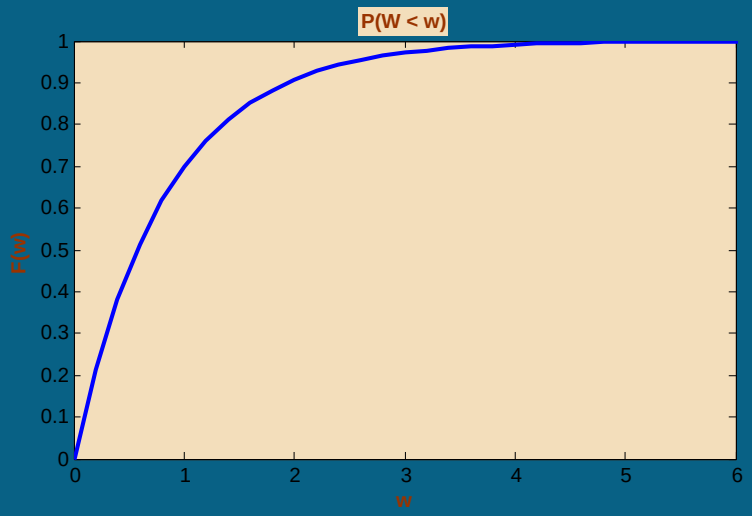
The posterior distribution of N

We used a large Monte Carlo sample to estimate the posterior density of N. There is only a very small probability that there are more than 15 clients in the system.



The posterior distribution of W

There is only a very small probability that a client spends over 5 minutes in the system.



Extensions

- Calculation of busy period and transient time distributions.

- Extension to other queueing systems:

  - ◇ Markovian systems with various or unlimited numbers of servers, e.g. Armero and Bayarri (1997).
  - ◇ Networks of queues. Armero and Bayarri (1999).
  - ◇ Non Markovian systems. See e.g. Wiper (1998), Ausín et al (2004).

References

Armero, C. and Bayarri, M.J. (1994a). Bayesian prediction in M/M/1 queues. *Queueing Systems*, **15**, 419-426.

Armero, C. and Bayarri, M.J. (1994b). Prior assessments for prediction in queues. *The Statistician*, **43**, 139-153

Armero, C. and Bayarri, M.J. (1997). Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, **58**, 241–261.

Armero, C. and Bayarri, M.J. (1999). Dealing with Uncertainties in Queues and Networks of Queues: A Bayesian Approach. In: *Multivariate Analysis, Design of Experiments and Survey Sampling*, Ed. S. Ghosh. New York: Marcel Dekker, pp 579–608.

Ausín, M.C., Wiper, M.P. and Lillo, R.E. (2004). Bayesian estimation for the M/G/1 queue using a phase type approximation. *Journal of Statistical Planning and Inference*, **118**, 83–101.

Gross, D. and Harris, C.M. (1985). *Fundamentals of Queueing Theory* (2nd ed.). New York: Wiley.

Hall, R.W. (1991). *Queueing Methods*. New Jersey: Prentice Hall.

Ruggeri, F., Wiper, M.P. and Rios Insua, D. (1996). Bayesian models for correlation in M/M/1 queues. *Quaderno IAMI 96.8*, CNR-IAMI, Milano.

Wiper, M.P. (1998). Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, **69**, 65–79.

4. Gaussian models

Box

Tiao

Objective

Introduce Bayesian inference for one and two sample problems with normally distributed data, illustrating the differences and similarities between Bayesian solutions and classical solutions.

Recommended reading

- Box, G.E. and Tiao, G.C. (1992). *Bayesian inference in statistical analysis*, Chapter 2.

- Lee, P.M. (2004). *Bayesian Statistics: An Introduction*, Chapter 2.

- Wiper, M.P., Girón, F.J. and Pewsey, A. (2008). Objective Bayesian inference for the half-normal and half-t distributions. *Communications in Statistics: Theory and Methods*, **37**, 3165–3185.

http://docubib.uc3m.es/WORKINGPAPERS/WS/ws054709.pdf

Introduction: one sample inference problems

Initially, we shall consider the problem of estimating the mean, μ, and/or variance, σ^2, of a single normal population, $X|\mu, \sigma^2 \sim \mathcal{N}\left(\mu, \sigma^2\right)$, with density

$$f\left(x|\mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

from a Bayesian perspective.

We shall consider three different scenarios:

a) inference for μ when σ^2 is known,

b) inference for σ^2 when μ is known,

c) joint inference for μ, σ^2.

and will assume throughout that we observe a sample, $\mathbf{x}$, of size n.

Inference for μ when σ is known

In order to develop a conjugate prior, we must first represent the normal distribution in exponential family form. We have:

$$f(x|\mu) = \underbrace{\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}_{C(\mu)} \underbrace{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)}_{h(x)} \exp\left(\underbrace{\frac{\mu}{\sigma^2}}_{\phi(\mu)} \underbrace{x}_{s(x)}\right)$$

which is a one-parameter exponential family representation as in Definition 3.

A conjugate prior distribution

Given this exponential family representation, from Theorem 9, we can derive the form of a conjugate prior distribution as

$$\begin{aligned} p(\mu) &\propto C(\mu)^a \exp(\phi(\mu)b) \\ &\propto \exp\left(-\frac{\mu^2}{2\sigma^2}\right)^a \exp\left(\frac{\mu}{\sigma^2}b\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left[a\mu^2 - 2b\mu\right]\right) \\ &\propto \exp\left(-\frac{a}{2\sigma^2}\left[\mu - \frac{b}{a}\right]^2\right) \end{aligned}$$

which is also a normal distribution, $\mu \sim \mathcal{N}\left(m, \tau^2\right)$, where $m = \frac{b}{a}$ and $\tau^2 = \frac{\sigma^2}{a}$.

The predictive and posterior distributions

Theorem 16
Let $X|\mu \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with prior distribution $\mu \sim \mathcal{N}\left(m, \tau^2\right)$. Then the predictive distribution of X is

$$X \sim \mathcal{N}\left(m, \sigma^2 + \tau^2\right).$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior distribution is $\mu|\mathbf{x} \sim \mathcal{N}\left(m^\star, \tau^{\star 2}\right)$ where

$$\begin{aligned} m^\star &= \frac{\frac{1}{\tau^2}m + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \\ \tau^{\star 2} &= \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}. \end{aligned}$$

Derivation of the predictive distribution

Proof We can write $X = \mu + \epsilon$ where $\mu \sim \mathcal{N}\left(m, \tau^2\right)$ and $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ are independent. Thus, the marginal distribution of X is

$$X \sim \mathcal{N}\left(m + 0, \sigma^2 + \tau^2\right)$$

and the result follows. ∎

We can derive the posterior distribution either directly via Bayes theorem, or alternatively via Theorem 9. In order to apply Bayes theorem, it is first useful to give a general expression for the normal likelihood function.

The normal likelihood function

Theorem 17

If a sample, $\mathbf{x}$, of size n is taken from the normal distribution $X|\mu, \sigma^2 \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then the likelihood function is

$$l(\mu, \sigma|\mathbf{x}) = (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\mu - \bar{x})^2\right]\right)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Proof

$$
\begin{aligned}
l(\mu, \sigma|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2\right) \\
&= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\mu - \bar{x})^2\right]\right)
\end{aligned}
$$

and the result follows. ∎

Now the posterior distribution can be derived from Bayes theorem in the usual way.

Derivation of the posterior distribution via Bayes Theorem

Proof

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)l(\mu|\mathbf{x}) \\
&\propto \exp\left(-\frac{1}{2\tau^2}(\mu-m)^2\right)\sigma^{-n}\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\mu-\bar{x})^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\left(\frac{\mu-m}{\tau}\right)^2+n\left(\frac{\mu-\bar{x}}{\sigma}\right)^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\mu^2-2\left(\frac{m}{\tau^2}+\frac{n\bar{x}}{\sigma^2}\right)\mu\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\left[\mu-\frac{\frac{m}{\tau^2}+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right]^2\right)
\end{aligned}
$$

which is a normal density; $\mu|\mathbf{x} \sim \mathcal{N}\left(\frac{\frac{m}{\tau^2}+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$. ∎

Derivation via Theorem 9

Proof In deriving the prior distribution, we found that $m = b/a$ and $\tau^2 = \sigma^2/a$ where $f(\mu) \propto C(\mu)^a \exp(\phi(\mu)b)$ is the conjugate representation of the prior in Theorem 9. Therefore, $a = \sigma^2/\tau^2$ and $b = m\sigma^2/\tau^2$. Also, we have $s(x) = x$ in our exponential family representation, which implies that, a posteriori,

$$\begin{aligned} a^\star &= a + n = \frac{\sigma^2}{\tau^2} + n = \sigma^2\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right) \\ b^\star &= b + \sum_{i=1}^{n} s(x_i) = m\frac{\sigma^2}{\tau^2} + n\bar{x} = \sigma^2\left(\frac{1}{\tau^2}m + \frac{n}{\sigma^2}\bar{x}\right), \end{aligned}$$

so that the posterior distribution is given by $\mu|\mathbf{x} \sim \mathcal{N}\left(m^\star, \tau^{\star 2}\right)$ where

$$m^\star = \frac{b^\star}{a^\star} = \frac{\frac{m}{\tau^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \qquad \tau^{\star 2} = \frac{\sigma^2}{a^\star} = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

How to remember the updating formulae

There is a simple way of remembering the normal updating formulae. Note firstly that the maximum likelihood estimator for μ is

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Now, the posterior *precision* is given by $\frac{1}{\tau^{\star 2}} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$ where $\frac{1}{\tau^2}$ is the prior precision and $\frac{n}{\sigma^2}$ is the precision of the estimator $\bar{X}$ so,

$$posterior\ precision = prior\ precision + precision\ of\ MLE.$$

Also, the posterior mean is simply a weighted average, $m^\star = wm + (1-w)\bar{x}$, where $w = \frac{1/\tau^2}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$ is proportional to the prior precision.

$$posterior\ mean = \frac{prior\ precision \times prior\ mean + precision\ of\ MLE \times MLE}{prior\ precision + precision\ of\ MLE}.$$

Interpretation

Write $\tau^2 = \sigma^2/\alpha$ so that the prior distribution is $\mu \sim \mathcal{N}\left(m, \frac{\sigma^2}{\alpha}\right)$ and the posterior distribution is

$$\mu|\mathbf{x} \sim \mathcal{N}\left(\frac{\alpha m + n\bar{x}}{\alpha + n}, \frac{\sigma^2}{\alpha + n}\right).$$

Thus we can interpret the information in the prior as being equivalent to the information in a sample of size $\alpha = \frac{\sigma^2}{\tau^2}$ with sample mean m.

Limiting results and comparison with classical results

Suppose that we let $\alpha \to 0$ (or equivalently, $\tau^2 \to \infty$). Then the posterior distribution tends to

$$\mu|\mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

and the posterior mean is equal to the MLE and, for example, a posterior 95% *credible interval*

$$\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$$

is (numerically) equal to the classical 95% confidence interval.

The limiting (improper) prior distribution in this case is uniform,

$$p(\mu) \propto 1.$$

Inference for σ when μ is known

In order to find a conjugate density for σ (or σ^2), we can again represent the normal density in one-dimensional exponential family form as

$$f(x|\sigma) = \underbrace{\frac{1}{\sigma}}_{C(\sigma)} \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \exp\left(\underbrace{-\frac{1}{2\sigma^2}}_{\phi(\sigma)} \underbrace{(x-\mu)^2}_{s(x)} \right).$$

This implies that a conjugate prior for σ takes the form

$$p(\sigma) \propto \sigma^{-c} \exp\left(-\frac{d}{2\sigma^2}\right)$$

for some values of c, d. This is not a well known distributional form, although we can derive a simple prior distribution using a transformation.

A conjugate prior for the precision ϕ

It is much easier to work in terms of the precision $\phi = \frac{1}{\sigma^2}$. In this case, we have $X|\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$ and

$$f(x|\phi) = \sqrt{\frac{\phi}{2\pi}} \exp\left(-\frac{\phi}{2}(x-\mu)^2\right)$$

which is already in exponential family form and implies that a conjugate prior density for ϕ is

$$p(\phi) \propto \left(\sqrt{\phi}\right)^c \exp\left(-\frac{\phi}{2}d\right)$$

for some c, d, which is a gamma density, $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$, where $a = c + 2$ and $b = d$.

Note that using the change of variables formula on the prior for σ, we could also show that the implied prior for ϕ has the same structure.

Inference for ϕ when μ is known

Theorem 18
Let $X|\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$ and suppose that $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$ for some a, b. Then, the marginal distribution of X is

$$\frac{X-\mu}{\sqrt{b/a}} \sim \mathcal{T}_a$$

and given a sample $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior distribution of ϕ is

$$\phi|\mathbf{x} \sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b+(n-1)s^2+n(\mu-\bar{x})^2}{2}\right).$$

Proof of the predictive distribution

Proof Define $Z = \sqrt{\phi}(X - \mu) \sim \mathcal{N}(0, 1)$ and $Y = b\phi \sim \chi^2_a$ and therefore, from standard distribution theory we know that $\frac{Z}{\sqrt{Y/a}} \sim \mathcal{T}_a$, is a Student's t distributed random variable with a degrees of freedom and transforming back, we have

$$\frac{Z}{\sqrt{Y/a}} = \frac{X - \mu}{\sqrt{b/a}}$$

which proves the predictive distribution formula. ∎

In order to demonstrate the formula for the posterior distribution, it is first convenient to give the formula for the reparameterized normal likelihood function.

The reparameterized normal likelihood function

Theorem 19
If a sample, $\mathbf{x}$, of size n is taken from the normal distribution $X| \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$, then the likelihood function is

$$l(\mu, \phi|\mathbf{x}) = (2\pi)^{-\frac{n}{2}}\phi^{\frac{n}{2}} \exp\left(-\frac{\phi}{2}\left[(n-1)s^2 + n(\mu - \bar{x})^2\right]\right)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Proof This follows immediately by substituting $\phi = 1/\sigma^2$ in Theorem 17. ∎

Given this form for the normal likelihood function, the posterior distribution in Theorem 18 can be derived immediately from Bayes Theorem by combining prior and likelihood.

Interpretation

In this case, we can interpret the information in the prior as equivalent to the information in a sample $\mathbf{y}$ of size a where the sufficient statistic $\sum_{i=1}^{n}(y_i - \mu)^2 = b$.

Limiting results

Letting $a, b \longrightarrow 0$, we have the limiting improper prior $f(\phi) \propto \frac{1}{\phi}$ and the limiting posterior distribution

$$\phi|\mathbf{x} \sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}\right) \quad \text{so that} \quad E[\phi|\mathbf{x}] = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{-1}$$

which is the classical MLE of ϕ. Note however that

$$E\left[\frac{1}{\phi}\middle|\mathbf{x}\right] = E\left[\sigma^2|\mathbf{x}\right] = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-2}$$

which is not the MLE of σ^2.

Interval estimation for σ (or σ^2)

Theorem 20

Let $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$ and define $\sigma = 1/\sqrt{\phi}$. Then the constant, c, such that $p = P(\sigma < c) = P(\sigma^2 < c^2)$, for given p, is $c = \sqrt{\frac{b}{\chi^2_a(1-p)}}$.

Proof

$$\begin{aligned} p &= P(\sigma < c) = P\left(\sigma^2 < c^2\right) = P\left(\phi > \frac{1}{c^2}\right) \\ &= P\left(b\phi > \frac{b}{c^2}\right) = P\left(Y > \frac{b}{c^2}\right) \quad \text{where } Y \sim \chi^2_a \\ \frac{b}{c^2} &= \chi^2_a(1-p) \\ c &= \sqrt{\frac{b}{\chi^2_a(1-p)}}. \end{aligned}$$

∎

Limiting case posterior interval estimation of σ and σ^2

Theorem 21

Let $X|\phi \sim \mathcal{N}(\mu, 1/\phi)$ with (improper) prior distribution $f(\phi) \propto 1/\phi$. Then

$$\phi|\mathbf{x} \sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}\right).$$

Moreover, let $\sigma = 1/\sqrt{\phi}$. Then a $100(1-p)\%$ credible interval for σ is

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_n\left(1-\frac{p}{2}\right)}} < \sigma < \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_n\left(\frac{p}{2}\right)}}$$

and a $100(1-p)\%$ credible interval for σ^2 is

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_n\left(1-\frac{p}{2}\right)} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_n\left(\frac{p}{2}\right)},$$

which are equal to the classical $100(1-p)\%$ confidence intervals for σ and σ^2.

Proof We have already given the posterior density for ϕ and the credible intervals follow immediately from Theorem 20 by setting $a = n$ and $b = \sum_{i=1}^{n}(x_i - \mu)^2$. ∎

Joint inference for μ and ϕ

When both parameters are unknown, it is again convenient to model in terms of the precision, ϕ, instead of the variance. Representing the normal density in exponential family form, we have

$$f(x|\mu,\phi) = \underbrace{\sqrt{\phi}\exp\left(-\frac{\phi\mu^2}{2}\right)}_{C(\mu,\phi)} \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)} \exp\left(\underbrace{\phi\mu}_{\phi_1(\mu,\phi)} \underbrace{x}_{s_1(x)} \underbrace{-\frac{\phi}{2}}_{\phi_2(\mu,\phi)} \underbrace{x^2}_{s_2(x)} \right)$$

so that a conjugate prior for μ, ϕ can be found as

$$\begin{aligned} p(\mu,\phi) &\propto \left(\sqrt{\phi}\exp\left(-\frac{\phi\mu^2}{2}\right)\right)^a \exp\left(b_1\phi\mu - b_2\frac{\phi}{2}\right) \quad \text{for some } a, b_1, b_2 \\ &\propto \phi^{\frac{a}{2}}\exp\left(-\frac{\phi}{2}\left[a\mu^2 - 2b_1\mu + b_2\right]\right) \propto \phi^{\frac{a}{2}}\exp\left(-\frac{\phi}{2}\left[d + a(\mu-c)^2\right]\right) \end{aligned}$$

where $c = b_1/a$ and $d = b_2 - b_1^2/a$. This is a *normal-gamma distribution*.

The normal-gamma distribution

Definition 7
Let $\mathbf{Y} = (Y_1, Y_2)$ and suppose that $Y_1|Y_2 \sim \mathcal{N}\left(\theta, \frac{1}{\lambda Y_2}\right)$ and $Y_2 \sim \mathcal{G}(\delta, \kappa)$. Then the distribution of $\mathbf{Y}$ is called a normal-gamma distribution, $\mathbf{Y} \sim \mathcal{NG}(\theta, \lambda, \delta, \kappa)$. The normal-gamma density function is

$$f(\mathbf{y}) = \frac{\kappa^\delta}{\Gamma(\delta)}\sqrt{\frac{\lambda}{2\pi}}y_2^{\frac{2\delta-1}{2}}\exp\left(-\frac{y_2}{2}\left[2\kappa + \lambda(y_1-\theta)^2\right]\right) \quad \text{for } -\infty < y_1 < \infty, y_2 > 0.$$

The following theorem gives the properties of the marginal distribution of Y_1.

Theorem 22
If $\mathbf{Y} = (Y_1, Y_2) \sim \mathcal{NG}(\theta, \lambda, \delta, \kappa)$, then $\frac{Y_1-\theta}{\sqrt{\kappa/(\lambda\delta)}} \sim \mathcal{T}_{2\delta}$. Furthermore, $E[Y_1] = \theta$ if $\delta > \frac{1}{2}$ and $V[Y_1] = \frac{\kappa}{\lambda(\delta-1)}$ if $\delta > 1$.

Proof

$$
\begin{aligned}
f(y_1) &= \int f(y_1, y_2)\, dy_2 \\
&= \frac{\kappa^\delta}{\Gamma(\delta)}\sqrt{\frac{\lambda}{2\pi}}\int_0^\infty y_2^{\frac{2\delta-1}{2}} \exp\left(-\frac{y_2}{2}\left[2\kappa + \lambda(y_1-\theta)^2\right]\right) dy_2 \\
&\propto \int_0^\infty y_2^{\frac{2\delta+1}{2}-1} \exp\left(-\frac{y_2}{2}\left[2\kappa + \lambda(y_1-\theta)^2\right]\right) dy_2 \\
&\propto \left(2\kappa + \lambda(y_1-\theta)^2\right)^{\frac{2\delta+1}{2}} \propto \left(1 + \frac{1}{2\delta}\left(\frac{y_1-\theta}{\sqrt{\kappa/(\lambda\delta)}}\right)^2\right)^{\frac{2\delta+1}{2}}
\end{aligned}
$$

and defining $T = \frac{Y_1-\theta}{\sqrt{\kappa/(\lambda\delta)}}$, we see that $f(t) \propto \left(1 + \frac{1}{2\delta}t^2\right)^{\frac{2\delta+1}{2}}$ which is the nucleus of a Student's t density with 2δ degrees of freedom. The mean and variance formulae follow from writing $Y_1 = \theta + \sqrt{\kappa/(\lambda\delta)}T$ and noting that $E[T] = 0$ for $2\delta > 1$ and $V[T] = \frac{2\delta}{2\delta-2} = \frac{\delta}{\delta-1}$ for $2\delta > 2$. ∎

Conjugate inference for μ, ϕ given a normal-gamma prior

Theorem 23
Let $X|\mu,\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$ and $\mu,\phi \sim \mathcal{NG}\left(m, \alpha, \frac{a}{2}, \frac{b}{2}\right)$ so that $\mu|\phi \sim \mathcal{N}\left(m, \frac{1}{\alpha\phi}\right)$ and $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$. Then the marginal distribution of X is

$$\frac{X-m}{\sqrt{(\alpha+1)b/(\alpha a)}} \sim \mathcal{T}_a.$$

Given a sample, $\mathbf{x}$, the posterior distribution of μ, ϕ is

$$\begin{aligned}
\mu,\phi &\sim \mathcal{NG}\left(m^\star, \alpha^\star, \frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
m^\star &= \frac{\alpha m + n\bar{x}}{\alpha+n} \\
\alpha^\star &= \alpha + n \\
a^\star &= a + n \\
b^\star &= b + (n-1)s^2 + \frac{\alpha n}{\alpha+n}(m-\bar{x})^2.
\end{aligned}$$

Proof Note first that

$$X|\phi \sim \mathcal{N}\left(m, \left(1+\frac{1}{\alpha}\right)\frac{1}{\phi}\right).$$

Therefore, X and ϕ have a normal gamma distribution.

$$X, \phi \sim \mathcal{NG}\left(m, \left(1+\frac{1}{\alpha}\right)^{-1}, a/2, b/2\right)$$

and the marginal distribution of X follows immediately from Theorem 22 setting $(\theta, \lambda, \delta, \kappa) = \left(m, \left(1+\frac{1}{\alpha}\right)^{-1}, a/2, b/2\right)$. In order to calculate the posterior distribution, we can use Bayes theorem:

$$
\begin{aligned}
p(\mu,\phi|\mathbf{x}) &\propto p(\mu,\phi)l(\mu,\phi|\mathbf{x}) \\
&\propto \phi^{\frac{a-1}{2}}\exp\left(-\frac{\phi}{2}\left[b+\alpha(\mu-m)^2\right]\right)\phi^{\frac{n}{2}}\exp\left(-\frac{\phi}{2}\left[(n-1)s^2+n(\mu-\bar{x})^2\right]\right) \\
&\propto \phi^{\frac{a+n-1}{2}}\exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+(\alpha+n)\mu^2-2(\alpha m+n\bar{x})\mu+\alpha m^2+n\bar{x}^2\right]\right) \\
&\propto \phi^{\frac{a+n-1}{2}}\exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+(\alpha+n)\left(\mu-\frac{\alpha m+n\bar{x}}{\alpha+n}\right)^2+\alpha m^2+n\bar{x}^2-\right.\right. \\
&\qquad \left.\left.-\frac{(\alpha m+n\bar{x})^2}{\alpha+n}\right]\right) \\
&\propto \phi^{\frac{a+n-1}{2}}\exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+\frac{\alpha n}{\alpha+n}(m-\bar{x})^2+\right.\right. \\
&\qquad \left.\left.(\alpha+n)\left(\mu-\frac{\alpha m+n\bar{x}}{\alpha+n}\right)^2\right]\right)
\end{aligned}
$$

which is the nucleus of the given normal-gamma density. ∎

Results with the limiting prior

Suppose that we use the improper prior $p(\mu, \phi) \propto \frac{1}{\phi}$. Then,

$$\begin{aligned} p(\mu, \phi|\mathbf{x}) &\propto \frac{1}{\phi}\phi^{\frac{n}{2}} \exp\left(-\frac{\phi}{2}\left[(n-1)s^2 + n(\mu - \bar{x})^2\right]\right) \\ &\propto \phi^{\frac{(n-1)-1}{2}} \exp\left(-\frac{\phi}{2}\left[(n-1)s^2 + n(\mu - \bar{x})^2\right]\right) \quad \text{so that} \\ \mu, \phi|\mathbf{x} &\sim \mathcal{NG}\left(\bar{x}, n, \frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \end{aligned}$$

Now, from Theorem 22, we have that

$$\frac{\mu - \bar{x}}{\sqrt{(n-1)s^2/(n(n-1))}} = \frac{\mu - \bar{x}}{s/\sqrt{n}} \sim \mathcal{T}_{n-1}.$$

Therefore, for example, a 100(1-p)% posterior credible interval for μ is

$$\bar{x} \pm \frac{s}{\sqrt{n}} \mathcal{T}_{n-1}\left(1 - \frac{p}{2}\right)$$

which is equal to the classical confidence interval.

Semi-conjugate inference via Gibbs sampling

The fully conjugate, normal-gamma prior distribution is slightly unnatural from the point of view of real prior choice, in that the distribution of μ depends on the unknown model precision ϕ. An alternative is to assume that μ and ϕ have *independent* prior distributions, say

$$\mu \sim \mathcal{N}\left(m, \frac{1}{\psi}\right) \quad \phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$$

where m, ψ, a, b are all known.

In this case, the joint posterior distribution given a sample, $\mathbf{x}$, is

$$p(\mu, \phi|\mathbf{x}) \propto \phi^{\frac{a+n}{2}-1} \exp\left(-\frac{1}{2}\left[\left(b + (n-1)s^2\right)\phi + n\phi(\mu - \bar{x})^2 + \psi(\mu - m)^2\right]\right).$$

The marginal distributions of μ and ϕ are

$$\begin{aligned} p(\mu|\mathbf{x}) &\propto \exp\left(-\frac{\psi}{2}(\mu-m)^2\right)\left(b+(n-1)s^2+n(\mu-\bar{x})^2\right)^{-\frac{a+n}{2}} \\ p(\phi|\mathbf{x}) &\propto \frac{\phi^{\frac{a+n}{2}-1}}{\sqrt{n\phi+\psi}}\exp\left(-\frac{\phi}{2}\left[b+(n-1)s^2+\frac{n\psi}{n\phi+\psi}(m-\bar{x})^2\right]\right). \end{aligned}$$

Neither the joint posterior density or the marginal posterior densities has a standard form. One possible solution is to use numerical integration techniques to estimate the integration constants, moments etc. of these densities.

Another possibility is to use a Monte Carlo approach. In this case, direct Monte Carlo sampling cannot easily be applied. However, an indirect approach is available.

Note firstly that the *conditional* posterior densities of $\mu|\phi, \mathbf{x}$ and of $\phi|\mu, \mathbf{x}$ are available in closed form. We have

$$\begin{aligned} \mu|\phi, \mathbf{x} &\sim \mathcal{N}\left(\frac{n\phi\bar{x} + \psi m}{n\phi + \psi}, \frac{1}{n\phi + \psi}\right) \\ \phi|\mu, \mathbf{x} &\sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b + (n-1)s^2 + n(\mu - \bar{x})^2}{2}\right) \end{aligned}$$

It is straightforward to sample from these distributions. Therefore, we can set up a *Gibbs sampler* to give an approximate Monte Carlo sample from the joint posterior distribution.

Aside: The (two variable) Gibbs sampler

Assume that we wish to sample from the density of $\mathbf{X} = (X_1, X_2)$. Suppose that the conditional densities, $p(\cdot|X_2)$ and $p(\cdot|X_1)$ are known. Then the (two variable) Gibbs sampler (Geman and Geman 1984) is a scheme for iteratively sampling these conditional densities as follows.

1. $t = 0$. Fix an initial value $x_1^{(0)}$.

2. Sample $x_2^{(t)} \sim p\left(x_2|x_1^{(t-1)}\right)$.

3. Sample $x_1^{(t)} \sim p\left(x_1|x_2^{(t)}\right)$.

4. Set $t = t + 1$.

5. Go to 2.

As t increases, it can be demonstrated that the sampled values approximate a Monte Carlo sample from the joint density, $p(\mathbf{X})$ and can be used in the same way as any other Monte Carlo sample.

The Gibbs sampler is an example of a Markov chain Monte Carlo (MCMC) algorithm and will be discussed in full in chapter 6.

Applying the Gibbs sampler to sampling $p(\mu, \phi|\mathbf{x})$

In our case, we can implement the Gibbs sampler as follows.

1. $t = 0$. Fix an initial value $\mu^{(0)}$.

2. Generate $\phi^{(t)} \sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b+(n-1)s^2+n\left(\mu^{(t-1)}-\bar{x}\right)^2}{2}\right)$.

3. Generate $\mu^{(t)} \sim \mathcal{N}\left(\frac{n\phi^{(t)}\bar{x}+\psi m}{n\phi^{(t)}+\psi}, \frac{1}{n\phi^{(t)}+\psi}\right)$

4. Set $t = t + 1$.

5. Go to 2.

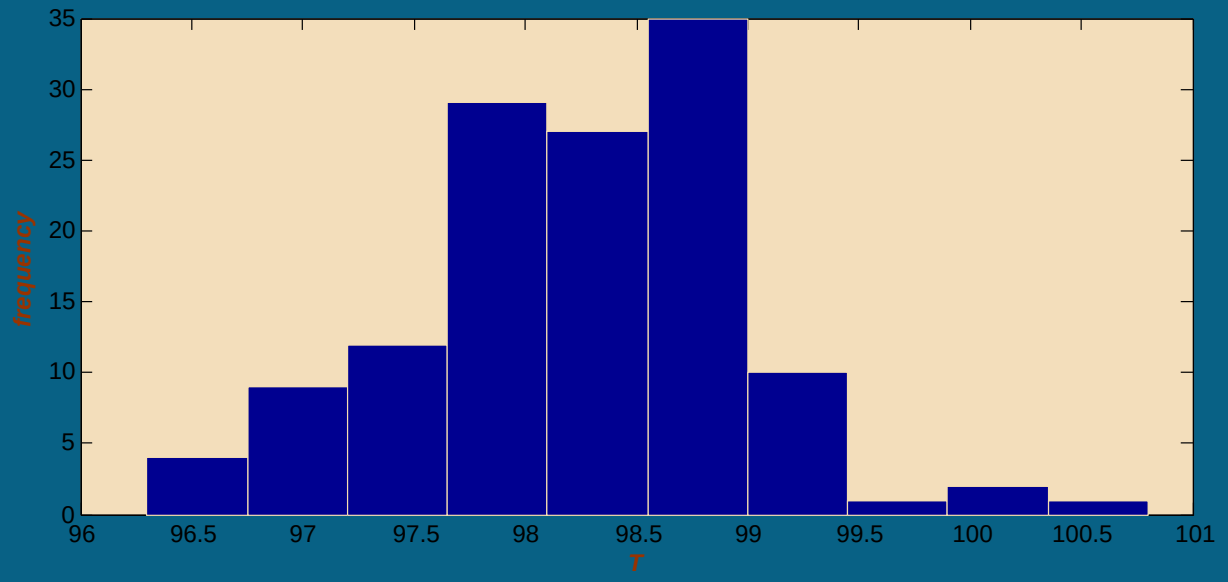
In order to estimate the marginal density of μ, for example, we can recall that $p(\mu|\mathbf{x}) = \int p(\mu|\phi, \mathbf{x})p(\phi|\mathbf{x}) \approx \frac{1}{T}\sum_{t=1}^{T} p\left(\mu|\phi^{(t)}, \mathbf{x}\right)$.

Normal body temperature example

Example 20

The normal core body temperature, T, of a healthy adult is supposed to be 98.6 degrees fahrenheit or 37 degrees celsius on average. As temperature can very around the mean, depending on given conditions, a normal model for temperatures, say $T|\mu, \phi \sim \mathcal{N}(\mu, 1/\phi)$, has been proposed.

Mackowiak et al (1992) measured the core body temperatures of 130 individuals and a histogram of the results is given below.



The sample mean temperature is $\bar{x} = 98.2492$ fahrenheit with sample standard deviation $s = 0.7332$. Thus, a classical 95% confidence interval for the true mean temperature is

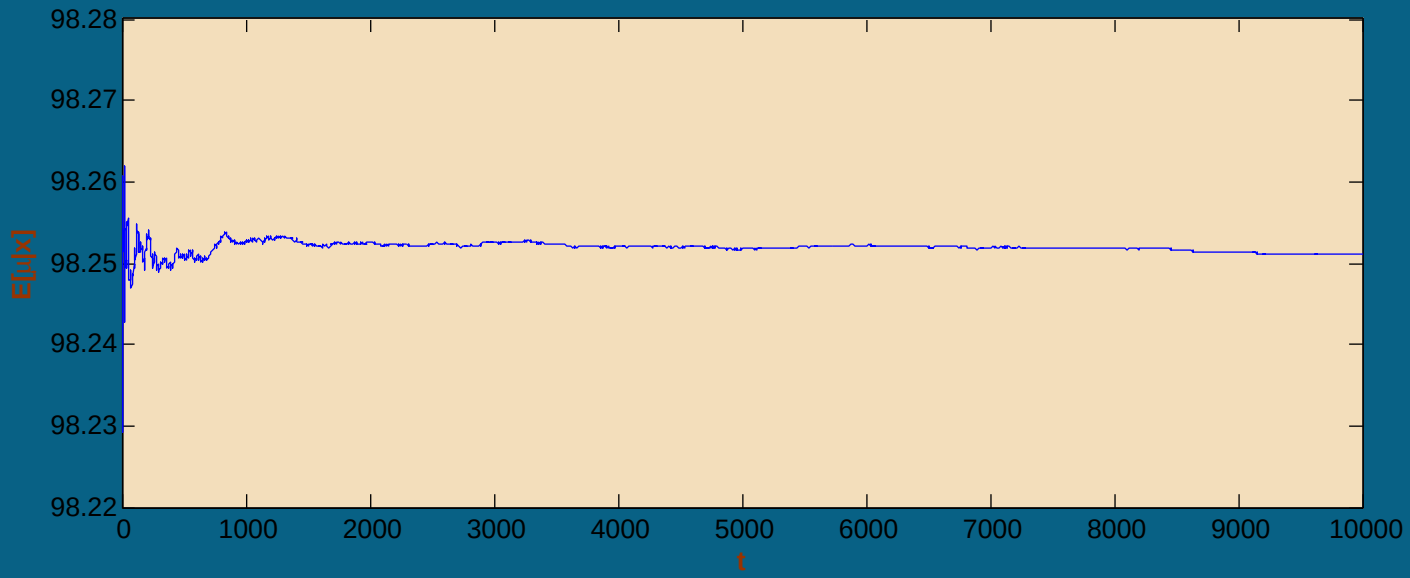
$$98.2492 \pm 1.96 \times 0.7332/\sqrt{130} = (98.1232, 98.3752)$$

and the hypothesis that the true temperature is 98.6 degrees is clearly rejected.

From a Bayesian viewpoint, we will analyze using two different prior distributions. Firstly, we assume the limiting prior $p(\mu, \phi) \propto \frac{1}{\phi}$ and secondly, we assume independent priors with relatively large variances, that is $\mu \sim \mathcal{N}(98.6, 1)$ and $\phi \sim \mathcal{G}\left(\frac{1}{2}, \frac{1}{2}\right)$.

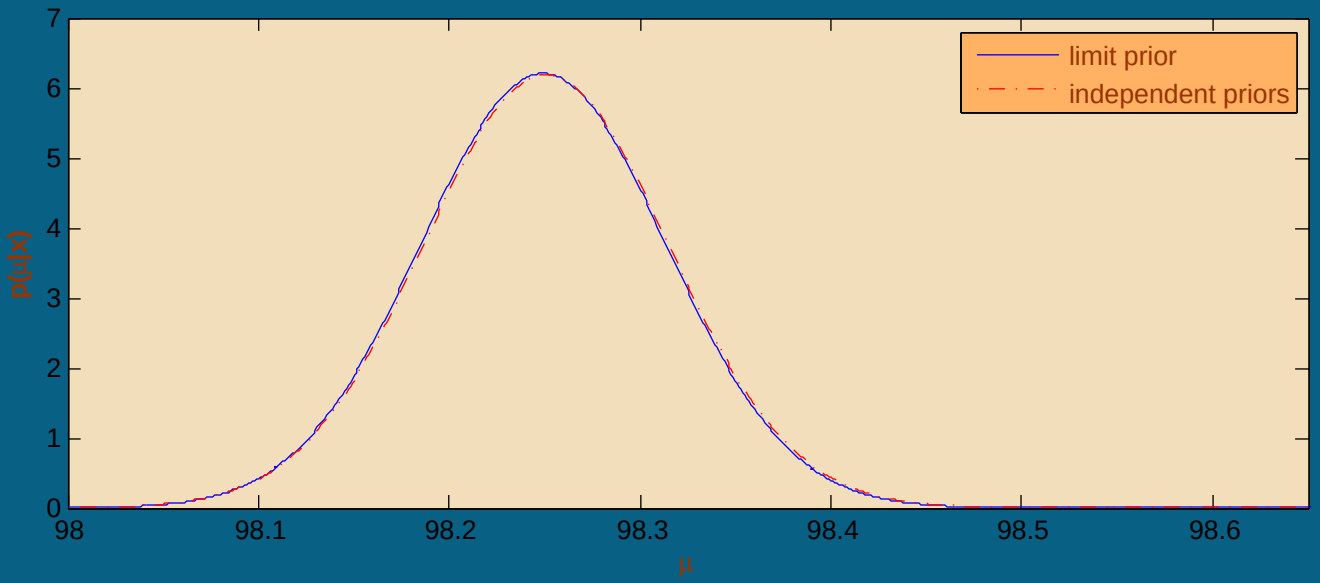
Given the limiting prior, we know that the posterior density of μ is $\frac{\mu - 98.2492}{0.7332/\sqrt{130}} \sim \mathcal{T}_{129}$ and a 95% posterior credible interval for μ coincides with the classical confidence interval.

Given the independent priors, the Gibbs sampler was run for 10000 iterations, starting at $\mu^{(0)} = 98.2492$. The following diagram shows a plot of the estimated posterior mean of μ versus the number of iterations.



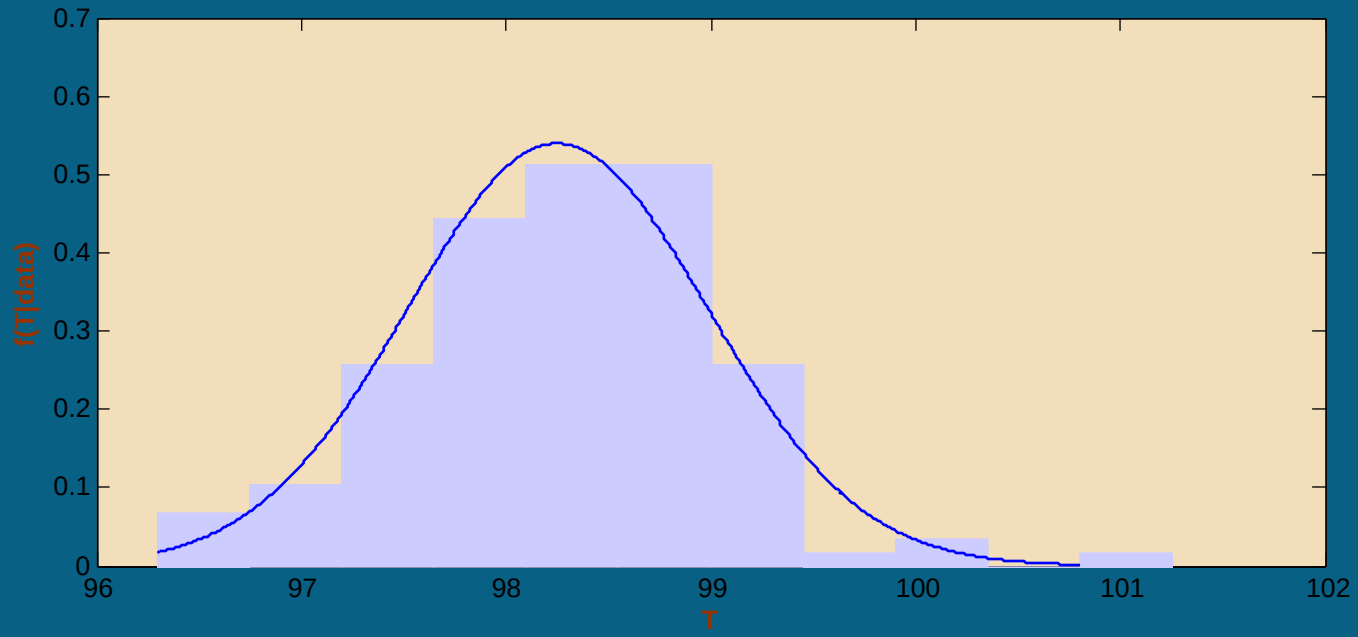
The estimated mean appears to have converged after 2000 iterations or less. The estimated posterior mean of μ was $E[\mu|\mathbf{x}] \approx 98.2512$.

The following diagram shows the posterior densities of μ given the limit and independent priors.



It can be seen that the posterior probability that μ is bigger than 98.6 fahrenheit is virtually zero.

The final diagram shows the predictive density of T given the observed data and the independent prior distributions.



The fit of the normal distribution is reasonable although it might be improved by incorporating other covariates, e.g. sex or heart rate.

A simple MATLAB code for executing the Gibbs sampler is as below.

```
n=length(temp); xbar=mean(temp); s=std(temp); % Summary statistics.
a=1; b=1; m=98.6; psi=1;                      % Priors for phi and mu.
iters=10000;mu=zeros(1,iters);phi=zeros(1,iters);                  % Sample
murange=98+[0:1000]/1000;pmu=zeros(1,length(murange));pmuuninf=pmu % initialization
xrange=min(temp)+[0:1000]*(max(temp)-min(temp))/1000; fx=zeros(1,length(xrange));


% Start of Gibbs sampler
mu(1)=xbar; phi(1)=gamrnd((a+n)/2,2/(b+(n-1)*s*s+n*((mu(1)-xbar)^2)));
for t=2:iters
    mu(t)=normrnd((psi*m+n*phi(t-1)*xbar)/(psi+n*phi(t-1)),1/sqrt(psi+n*phi(t-1)));
    phi(t)=gamrnd((a+n)/2,2/(b+(n-1)*s*s+n*((mu(t)-xbar)^2)));
    pmu=pmu+normpdf(murange,(psi*m+n*phi(t)*xbar)/(psi+n*phi(t)),1/sqrt(psi+n*phi(t)));
    fx=fx+normpdf(xrange,mu(t),1/sqrt(phi(t)));
end
pmu=pmu/iters; fx=fx/iters;  %Normalization.


pmuuninf=normpdf(murange,xbar,s/sqrt(n)); %Posterior density given 1/phi prior.
% Plots
figure; plot([1:iters],cumsum(mu)./[1:iters]); figure;
plot(murange,pmuuninf); hold on; plot(murange,pmu,'-.r'); hold off;
figure; histadjusted(temp,10); hold on; plot(xrange,fx); hold off;
```

Two sample problems

Various two sample problems have been studied.

1. Paired data samples: difference in means.

2. Unpaired samples:

   (a) Difference of two population means: variances known,
   (b) Difference of two population means: variances unknown but equal,
   (c) Difference of two population means: unknown variances,
   (d) Ratio of two population variances.

We shall consider Bayesian inference for each situation in turn.

Paired data

As in classical inference, by considering the differences of each pair, this reduces to a single normal sample inference problem.

Unpaired data: known variances

Assume that we have $X|\mu_X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ where σ_X^2 and σ_Y^2 are known. Suppose also that we use independent, uniform priors for μ_X and μ_Y. Then, given samples $\mathbf{x}$ and $\mathbf{y}$ of sizes n_X and n_Y respectively, we know that the posterior distributions are independent normal

$$\mu_X|\mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{\sigma_X^2}{n_X}\right) \quad \mu_Y|\mathbf{y} \sim \mathcal{N}\left(\bar{y}, \frac{\sigma_Y^2}{n_Y}\right).$$

Therefore, if we define $\delta = \mu_X - \mu_Y$ to be the difference in the two population means, then the posterior distribution of δ is

$$\delta|\mathbf{x}, \mathbf{y} \sim \mathcal{N}\left(\bar{x} - \bar{y}, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right).$$

In this case, a 95% credible interval for δ is

$$\bar{x} - \bar{y} \pm 1.96\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

which is equal to the corresponding classical confidence interval.

It is, of course, straightforward to generalize this result to the case when proper priors for μ_X and μ_Y are used.

Unpaired data: unknown, equal variances

Suppose now that $X|\mu_X, \phi \sim \mathcal{N}\left(\mu_X, \frac{1}{\phi}\right)$ and $Y|\mu_Y, \phi \sim \mathcal{N}\left(\mu_Y, \frac{1}{\phi}\right)$ have common, unknown precision ϕ and that we observe samples $\mathbf{x}$ and $\mathbf{y}$ of sizes n_X and n_Y respectively. Consider the joint prior distribution

$$p(\mu_X, \mu_Y, \phi) \propto \frac{1}{\phi}.$$

Then, the joint posterior distribution is

$$p(\mu_X, \mu_Y, \phi|\mathbf{x}, \mathbf{y}) \propto \phi^{\frac{n_X+n_Y}{2}-1} \exp\left(-\frac{\phi}{2}\left[(n_X-1)s_X^2 + (n_Y-1)s_Y^2 + n_X(\mu_X-\bar{x})^2 + n_Y(\mu_Y-\bar{y})^2\right]\right).$$

Given this joint distribution, we can see that

$$\mu_X|\phi, \mathbf{x}, \mathbf{y} \sim \mathcal{N}\left(\bar{x}, \frac{1}{n_X\phi}\right) \quad \mu_Y|\phi, \mathbf{x}, \mathbf{y} \sim \mathcal{N}\left(\bar{y}, \frac{1}{n_Y\phi}\right).$$

Integrating out μ_X and μ_Y successively from the joint distribution, we have

$$\phi|\mathbf{x} \sim \mathcal{G}\left(\frac{n_X + n_Y - 2}{2}, \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{2}\right).$$

Also, from the previous slide, if $\delta = \mu_X - \mu_Y$, then

$$\delta|\phi, \mathbf{x}, \mathbf{y} \sim \mathcal{N}\left(\bar{x} - \bar{y}, \frac{1}{\phi}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right).$$

Thus, the joint posterior distribution of δ and ϕ is normal-gamma

$$\delta, \phi|\mathbf{x}, \mathbf{y} \sim \mathcal{NG}\left(\bar{x} - \bar{y}, \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)^{-1}, \frac{n_X + n_Y - 2}{2}, \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{2}\right).$$

Therefore, the marginal posterior distribution of δ is Student's t:

$$\frac{\delta - (\bar{x} - \bar{y})}{\sqrt{\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)^{-1}(n_X+n_y-2)}}} = \frac{\delta - (\bar{x} - \bar{y})}{\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_y-2}}} \sim \mathcal{T}_{n_X+n_Y-2}.$$

Now, a 95% credible interval for δ is given by

$$\bar{x} - \bar{y} \pm \mathcal{T}_{n_X+n_y-2}(0.975)\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}s_c$$

where s_c^2 is the (classical) combined variance estimator

$$s_c^2 = \frac{(n_X-1)s_X^2+(n_Y-1)s_Y^2}{n_X+n_y-2}.$$

This is equal to the usual, classical confidence interval.

The Behrens-Fisher problem

Now assume that $X|\mu_X, \phi_X \sim \mathcal{N}\left(\mu_X, \frac{1}{\phi_X}\right)$ and $Y|\mu_Y, \phi_Y \sim \mathcal{N}\left(\mu_Y, \frac{1}{\phi_Y}\right)$ where all parameters are unknown. Given the (improper) prior distributions

$$p(\mu_X, \phi_X) \propto \frac{1}{\phi_X} \quad \text{and} \quad p(\mu_Y, \phi_Y) \propto \frac{1}{\phi_Y}$$

then we know that the marginal posterior distributions of μ_X and μ_Y are independent, shifted, scaled t distributions:

$$\frac{\mu_X - \bar{x}}{s_X/\sqrt{n_X}} \sim \mathcal{T}_{n_X-1} \quad \frac{\mu_Y - \bar{y}}{s_Y/\sqrt{n_Y}} \sim \mathcal{T}_{n_Y-1}.$$

Thus, if $\delta = \mu_X - \mu_Y$, we have

$$\delta = T_{n_X-1}\frac{s_X}{\sqrt{n_X}} - T_{n_Y-1}\frac{s_Y}{\sqrt{n_Y}} - (\bar{x} - \bar{y})$$

where T_d represents a t distributed variable with d degrees of freedom.

We could try to derive the density formula of δ directly from this expression. However, it is more straightforward to work in terms of a transformed variable.

Theorem 24
Let $\delta' = \frac{\delta-(\bar{x}-\bar{y})}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}}$. Then we can write

$$\delta' = T_{n_X-1}\sin\omega - T_{n_Y-1}\cos\omega$$

where $\omega = \tan^{-1}\frac{s_X/\sqrt{n_X}}{s_Y/\sqrt{n_Y}}$.

Proof From the previous page, we have

$$\begin{aligned}
\delta &= T_{n_X-1}\frac{s_X}{\sqrt{n_X}} - T_{n_Y-1}\frac{s_Y}{\sqrt{n_Y}} - (\bar{x}-\bar{y}) \\
\frac{\delta-(\bar{x}-\bar{y})}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}} &= \frac{T_{n_X-1}\frac{s_X}{\sqrt{n_X}}}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}} - \frac{T_{n_Y-1}\frac{s_Y}{\sqrt{n_Y}}}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}}
\end{aligned}$$

and the result follows by observing that the squares of the coefficients on the right hand side sum to 1. ∎

The Behrens-Fisher distribution

Definition 8
The distribution of

$$Y = T_{\nu_1} \sin\omega - T_{\nu_2} \cos\omega$$

where T_{ν_1} and T_{ν_2} are standard Student's t variables with ν_1 and ν_2 degrees of freedom, is the *Behrens-Fisher distribution*, (Behrens 1929) with parameters ν_1, ν_2 and ω. In this case, we write $Y \sim \mathcal{BF}(\nu_1, \nu_2, \omega)$.

Thus, from Theorem 24, we have

$$\delta' | \mathbf{x}, \mathbf{y} \sim \mathcal{BF}\left(n_X - 1, n_Y - 1 \tan^{-1} \frac{s_X / \sqrt{n_X}}{s_Y / \sqrt{n_Y}}\right).$$

Note that this Bayesian solution to the Behrens-Fisher problem corresponds to the *fiducial* solution and was first developed in this context by Fisher (1935).

Patil's approximation and other appoximate methods

Tables of the percentage points of the Behrens-Fisher distribution are available for given values of ω, see e.g. Kim and Cohen (1996). For other values, Patil's (1965) approximation may be applied.

If $Y \sim \mathcal{BF}(\nu_1, \nu_2, \omega)$ then we have $Y/a \approx \mathcal{T}_b$ where

$$
\begin{aligned}
b &= 4 + \left(\frac{\nu_1 \cos^2 \omega}{\nu_1 - 2} + \frac{\nu_2 \sin^2 \omega}{\nu_2 - 2} \right)^2 \left[\frac{\nu_1^2 \cos^4 \omega}{(\nu_1 - 2)^2 (\nu_1 - 4)} + \frac{\nu_2^2 \sin^4 \omega}{(\nu_2 - 2)^2 (\nu_2 - 4)} \right]^{-1} \\
a &= \frac{b}{b - 2} \left[\frac{\nu_1 \cos^2 \omega}{\nu_1 - 2} + \frac{\nu_2 \sin^2 \omega}{\nu_2 - 2} \right]^{-1}
\end{aligned}
$$

Alternative approximation methods are given by Ghosh (1975) or Willink (2004).

A Monte Carlo approach

Another approach is to use Monte Carlo sampling to approximate a Behrens Fisher distribution as follows. Suppose that $Y \sim \mathcal{BF}(\nu_1, \nu_2, \omega)$. Then we can generate a sample from the Behrens Fisher distribution as follows:

1. Fix some large M.

2. For $i = 1, \ldots, M$

   (a) Generate $t_1^{(i)} \sim \mathcal{T}_{\nu_1}$ and $t_2^{(i)} \sim \mathcal{T}_{\nu_2}$.
   (b) Define $y^{(i)} = t_1^{(i)} \cos\omega - t_2^{(i)} \sin\omega$.

Then, approximate 95% credible intervals for Y could be estimated using the sample quantiles. If M is increased sufficiently, the accuracy of these estimates can be derived to any fixed precision.

The frequentist approach to the Behrens-Fisher problem

No equivalent solution exists in frequentist inference because in this case, the sampling distribution of δ' depends upon the ratio, ϕ_X/ϕ_Y, of the unknown precisions.

The usual frequentist approximation to this sampling distribution is:

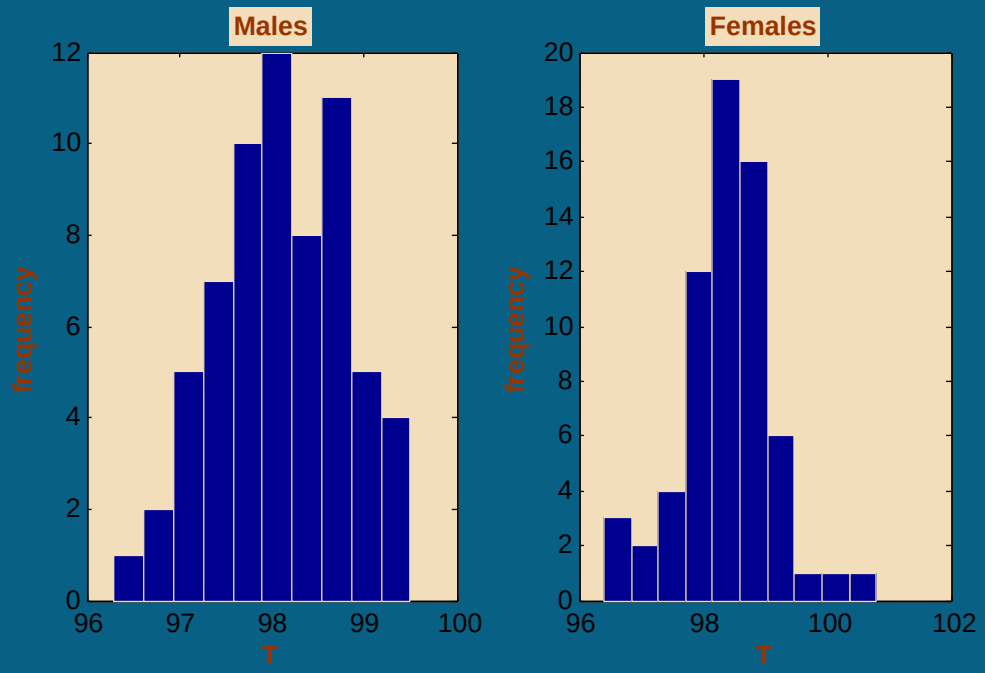
$$\delta' \approx \mathcal{T}_r \quad \text{where} \quad r = \frac{(s_X^2/n_X + s_Y^2/n_Y)^2}{\left(\frac{s_X^4/n_X^2}{n_X - 1} + \frac{s_Y^4/n_Y^2}{n_Y - 1}\right)}$$

However, we cannot know how good this approximation is for any given sample.

Another look at the normal body temperature example

Example 21

The sample contained 122 males and 22 females. The following two histograms illustrate the temperature distributions for both groups.



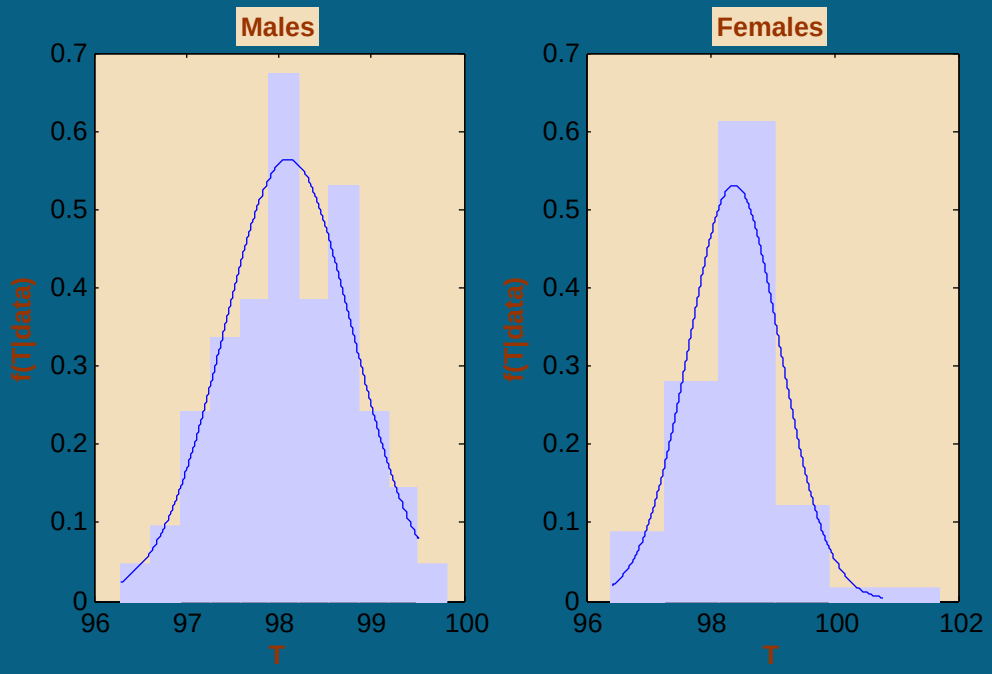
We may wish to question whether or not the mean body temperatures are the same in both groups.

The means (and standard deviations) for males and females are 98.1046 (0.6988) and 98.3938 (0.7435) respectively.

In this case, an approximate classical 95% confidence interval for the mean difference δ is given by $(-0.5396, -0.0388)$ and there is strong evidence that the mean body temperature is higher in females than in males.

Using a Bayesian approach with improper priors as outlined earlier, a 95% posterior credible interval for δ is given by $(-0.5359, -0.0426)$ which is slightly narrower than the classical interval.

Also, the fitted predictive posterior densities for both groups are given in the following diagram. The normal fits appear to have improved somewhat.



Comparing two population variances

Suppose as earlier that $X|\mu_X, \phi_X \sim \mathcal{N}\left(\mu_X, \frac{1}{\phi_X}\right)$ and $Y|\mu_Y, \phi_Y \sim \mathcal{N}\left(\mu_Y, \frac{1}{\phi_Y}\right)$ with the usual improper priors;

$$p(\mu_X, \phi_X) \propto \frac{1}{\phi_X} \quad p(\mu_Y, \phi_Y) \propto \frac{1}{\phi_Y}.$$

Then, from earlier, given samples of sizes n_X and n_Y, we know that

$$\phi_X|\mathbf{x} \sim \mathcal{G}\left(\frac{n_X - 1}{2}, \frac{(n_X - 1)s_X^2}{2}\right) \quad \phi_Y|\mathbf{y} \sim \mathcal{G}\left(\frac{n_Y - 1}{2}, \frac{(n_Y - 1)s_Y^2}{2}\right)$$

and therefore,

$$(n_X - 1)s_X^2\phi_X \sim \chi^2_{n_X-1} \quad (n_Y - 1)s_Y^2\phi_Y \sim \chi^2_{n_Y-1}.$$

Recalling that the ratio of two chi-squared distributions divided by their degrees of freedom is F distributed, we thus have

$$\frac{s_X^2}{s_Y^2}\frac{\phi_X}{\phi_Y} \sim \mathcal{F}_{n_Y-1}^{n_X-1}$$

and Bayesian and classical credible intervals will coincide.

Application II: Bayesian inference for the half-normal distribution

We will follow the analysis of Wiper et al (2008).

The half-normal distribution

If $Z \sim \mathcal{N}(0,1)$ then $X = |Z|$ has a half-normal distribution, $X \sim \mathcal{HN}(0,1)$, with probability density function

$$\begin{aligned} f(x) &= 2\phi(x) \quad \text{for } x > 0, \text{ where} \\ \phi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \quad \text{is the standard normal pdf.} \end{aligned}$$

The generalized half-normal distribution

In this case, we allow the location and scale parameters to vary. Thus, let $X = \xi + \eta|Z|$ where Z is a standard normal random variable as earlier. Then we say that X has a generalized half-normal distribution, $X|\xi, \eta \sim \mathcal{HN}(\xi, \eta)$, with density

$$f(x|\xi, \eta) = \frac{2}{\eta}\phi\left(\frac{x-\xi}{\eta}\right) \quad \text{for } x > \xi.$$

Applications

The half-normal distribution is a model for truncated data with applications in various areas:

- measuring body fat indices in athletes (Pewsey 2002, 2004).
- stochastic frontier modeling (Aigner et al 1977, Meeusen and van den Broeck, 1977).
- fluctuating asymmetry modeling (Palmer and Strobeck 1986).
- fibre buckling (Haberle 1991).
- blowfly dispersion (Dobzhansky and Wright 1947).

Classical inference for the half-normal distribution

Suppose we have a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from $\mathcal{HN}(\xi, \eta)$ and that we wish to estimate ξ (and η). Now from Pewsey (2002,2004):

- The MLE for ξ is $\hat{\xi} = x_{(1)}$, the minimum of the data.
- Uncorrected and bias corrected estimators for η are

$$\hat{\eta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_{(1)})^2} \qquad \hat{\eta}_{BC} = \sqrt{\frac{n}{n-1}}\hat{\eta}$$

- An (asymptotic) 95% confidence interval for ξ:

$$x_{(1)} + \log(\alpha)\hat{\eta}\Phi^{-1}\left(\frac{1}{2} + \frac{1}{2n}\right) < \xi < x_{(1)}.$$

Bayesian inference for the half-normal distribution

Inference is similar to the usual normal case. First we reparameterize by defining $\tau = \frac{1}{\eta^2}$ so that

$$f(x|\xi,\tau) = 2\sqrt{\tau}\phi\left(\sqrt{\tau}(x-\xi)\right).$$

Now we can define a generalized version of the normal-gamma prior distribution, that is the *right-truncated normal-gamma* distribution.

The right-truncated normal-gamma distribution

We say that, ξ, τ have a RTNG with parameters ξ_0, m, α, a, b,

$$\xi, \tau \sim \mathcal{RTNG}\left(\xi_0, m, \alpha, \frac{a}{2}, \frac{b}{2}\right) \quad \text{if}$$

$$p(\xi, \tau) = \frac{1}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \sqrt{\frac{\alpha}{2\pi}} \tau^{\frac{a+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[b + \alpha(\xi - m)^2\right]\right\}$$

for $\xi < \xi_0$ where $\Phi_d(\cdot)$ is the Student's t cdf with d degrees of freedom;

$$\Phi_d(z) = \int_{-\infty}^{z} \phi_d(y)\, dy \quad \text{and}$$

$$\phi_d(y) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \frac{1}{\sqrt{\pi d}} \left(1 + \frac{y^2}{d}\right)^{-\frac{d+1}{2}}$$

What is the motivation for this density?

$$p(\xi,\tau) = \frac{1}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \sqrt{\frac{\alpha}{2\pi}} \tau^{\frac{a+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[b + \alpha(\xi - m)^2\right]\right\}$$

$$p(\xi,\tau) = \frac{1}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)} \underbrace{\frac{(b/2)^{a/2}}{\Gamma(a/2)} \sqrt{\frac{\alpha}{2\pi}} \tau^{\frac{a+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[b + \alpha(\xi - m)^2\right]\right\}}_{\text{the usual normal gamma density}}$$

$$p(\xi,\tau) = \underbrace{\frac{1}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)}}_{\text{from the truncation}} \frac{(b/2)^{a/2}}{\Gamma(a/2)} \sqrt{\frac{\alpha}{2\pi}} \tau^{\frac{a+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[b + \alpha(\xi - m)^2\right]\right\}$$

The marginal density of ξ

$$
\begin{aligned}
p(\xi) &= \frac{1}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \frac{\Gamma\left(\frac{a+1}{2}\right)}{\Gamma\left(\frac{a}{2}\right)} \sqrt{\frac{\alpha}{b\pi}} \left(1 + \frac{1}{a}\left(\frac{\xi - m}{\sqrt{b/(a\alpha)}}\right)^2\right)^{-\frac{a+1}{2}} \\
&= \sqrt{\frac{a\alpha}{b}} \frac{\phi_a\left(\frac{\xi - m}{\sqrt{b/(a\alpha)}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \quad \text{for } \xi < \xi_0
\end{aligned}
$$

Thus, $\xi' = \frac{\xi - m}{\sqrt{b/(a\alpha)}} \sim \mathcal{T}_a$, truncated onto the region $\xi' < \frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}$.

Moments of ξ

$$E[\xi] = m - \sqrt{\frac{b}{a\alpha}} \frac{a + \left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)^2}{a-1} \frac{\phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \quad \text{for } a > 1$$

$$V[\xi] = \frac{b}{a\alpha(a-2)} \left\{ a - \left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right) \left(a + \left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)^2\right) \frac{\phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \right\}$$

$$- \frac{b}{a\alpha} \left(\frac{a + \left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)^2}{a-1} \frac{\phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \right)^2 \quad \text{for } a > 2$$

We have explicit formulae for all moments of ξ.

The conditional and marginal densities of τ

The conditional density of τ given ξ is the same as in the usual normal-gamma model, that is

$$\tau|\xi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b+\alpha(\xi-m)^2}{2}\right) \quad \text{for } \tau > 0.$$

The marginal distribution of τ is a *Gaussian modulated gamma distribution*, that is $\tau \sim \mathcal{GMG}(\sqrt{\alpha}(\xi_0 - m), a, b)$, with density

$$p(\tau) = \frac{\Phi\left(\sqrt{\alpha\tau}(\xi_0 - m)\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(a\alpha)}}\right)} \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \tau^{\frac{a}{2}-1} e^{-\frac{b}{2}\tau}.$$

Moments of τ

$$
\begin{aligned}
E[\tau] &= \frac{\Phi_{a+2}\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha(a+2))}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)} \frac{a}{b} \\
E\left[\tau^2\right] &= \frac{\Phi_{a+4}\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha(a+4))}}\right)}{\Phi_a\left(\frac{\xi_0 - m}{\sqrt{b/(\alpha a)}}\right)} \frac{a(a+2)}{b^2}
\end{aligned}
$$

We can derive all positive and negative moments of τ and therefore η.

The RTNG distribution is conjugate to the truncated normal

From standard normal gamma distribution theory, it is clear that given a sample $\mathbf{x}$ of half-normal data, and a RTNG prior, $\xi, \tau \sim \mathcal{RTNG}\left(\xi_0, m, \alpha, \frac{a}{2}, \frac{b}{2}\right)$, then the posterior distribution is

$$
\begin{aligned}
\xi, \tau | \mathbf{x} &\sim \mathcal{RTNG}\left(\xi_0^\star, m^\star, \alpha^\star, \frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
\xi_0^\star &= \min\{\xi_0, \mathbf{x}\} \\
m^\star &= \frac{\alpha m + n\bar{x}}{\alpha + n} \\
\alpha^\star &= \alpha + n \\
a^\star &= a + n \\
b^\star &= b + (n-1)s^2 + \frac{\alpha n}{\alpha + n}(m - \bar{x})^2
\end{aligned}
$$

The limiting case posterior

Suppose that we define the improper prior $p(\xi, \tau) \propto \frac{1}{\tau}$ for $-\infty < \xi < \infty$. Then, the posterior distribution is

$$\xi|\mathbf{x} \sim \mathcal{RTNG}\left(\min\{\mathbf{x}\}, \bar{x}, n, \frac{n-1}{2}, \frac{(n-1)s^2}{2}\right).$$

Also, the posterior mean in this case is:

$$E[\xi|\mathbf{x}] = \bar{x} - \frac{s}{\sqrt{n}} \frac{n + \left(\frac{\min\{\mathbf{x}\}-\bar{x}}{s/\sqrt{n}}\right)^2}{n-1} \frac{\phi_n\left(\frac{\min\{\mathbf{x}\}-\bar{x}}{s/\sqrt{n}}\right)}{\Phi_n\left(\frac{\min\{\mathbf{x}\}-\bar{x}}{s/\sqrt{n}}\right)} \neq \hat{\xi} = \min\{\mathbf{x}\}.$$

The classical and Bayesian estimator for ξ are different.

Comparison of the Bayesian posterior mean and the MLE

Bias

We simulated 1000 samples of various sizes from $\mathcal{HN}(0,1)$. The Table compares the estimated *biases* of the Bayesian posterior mean estimator of ξ with the MLE.

n	$\hat{\xi}$	$E[\xi\|\mathbf{X}]$
5	0.217	-0.063
10	0.118	-0.011
20	0.060	-0.003
50	0.025	-0.0005
100	0.012	-0.0002
1000	0.001	2×10^{-6}

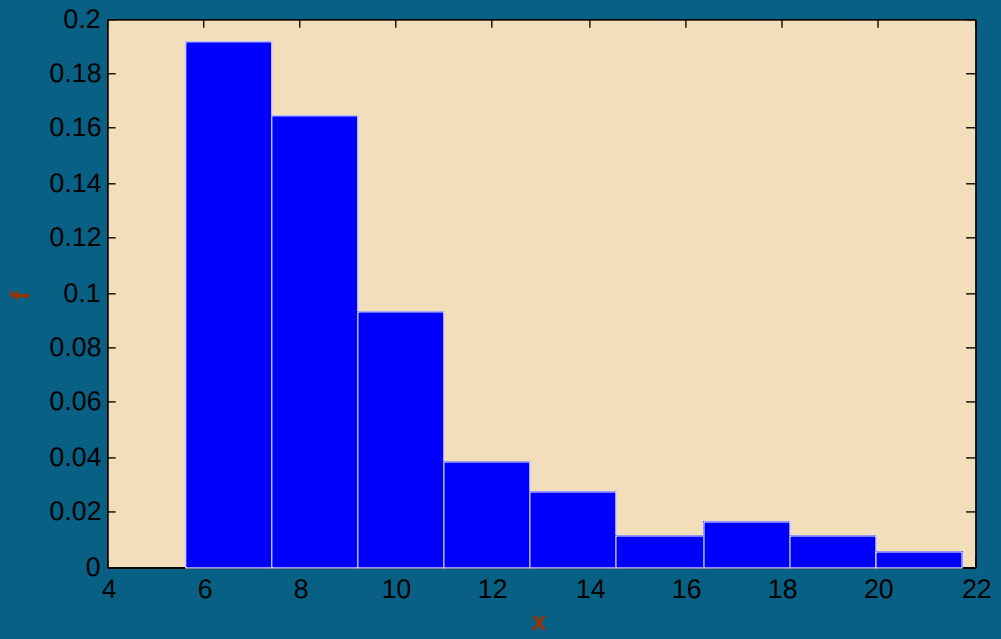
The bias of the Bayesian estimator is much lower than that of the MLE and compares with Pewsey's (2002,2004) unbiased estimator

Coverage of credible and confidence intervals

	Classical intervals				Bayesian intervals	
	Uncorrected		Bias Corrected			
n	coverage	length	coverage	length	coverage	length
5	0.880	0.585	0.904	0.655	0.950	0.910
10	0.916	0.332	0.927	0.350	0.945	0.398
20	0.937	0.177	0.943	0.181	0.951	0.192
50	0.944	0.073	0.946	0.074	0.949	0.076
100	0.948	0.037	0.948	0.037	0.950	0.038
1000	0.950	0.004	0.950	0.004	0.949	0.004

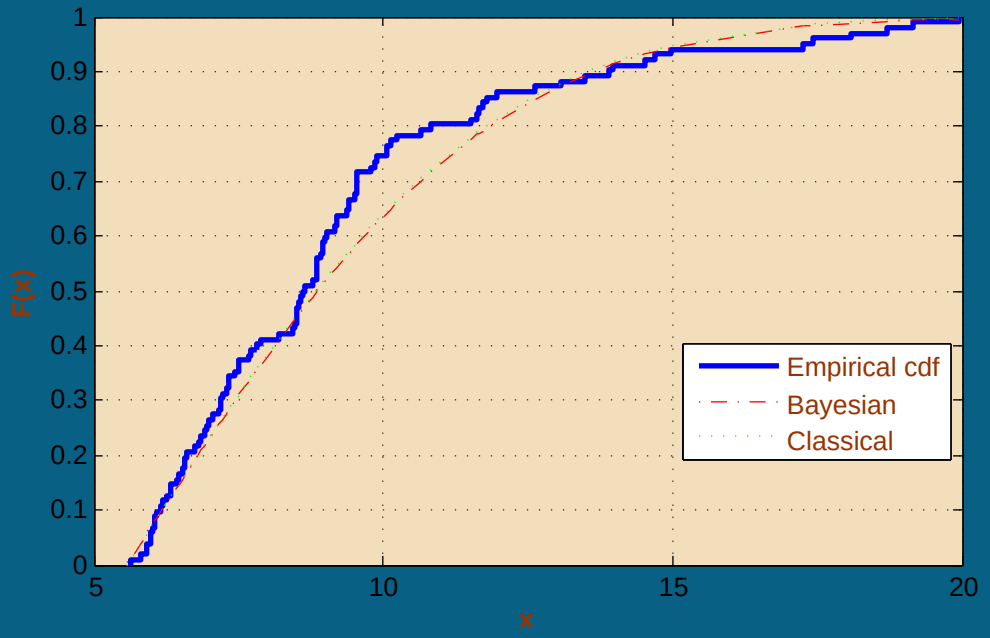
The athletes data example

Pewsey (2002,2004) analyzes data on the body fat indices of athletes.



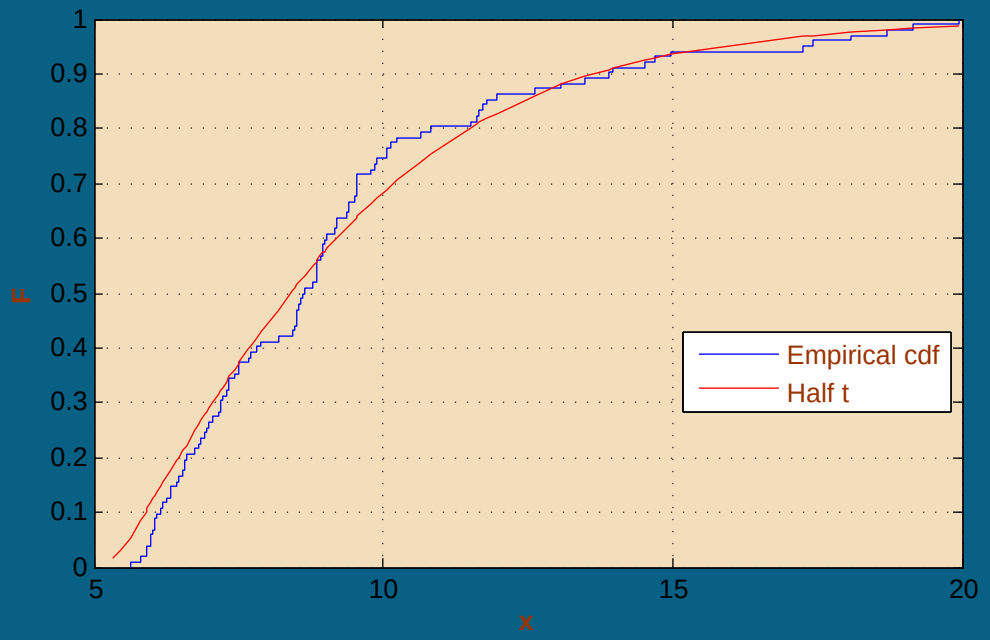
Maybe they can be well fitted by a half-normal model.

The predictive distribution function of X using Bayesian and classical methods



The fit doesn't look too good, but ...

Extension: the half-t model



. . . results of fitting a half-t distribution model are much more promising. But that's another story.

References

Aigner, D.J., Lovell, C.A.K., Schmidt, P. (1977). Formulation and estimation of stochastic frontier production models. *Journal of Econometrics*, **6**, 21–37.

Behrens, W.-V. (1929). Ein Beitrag zur Fehlerberchnung bei Wenigen Beobachtungen, *Landw. Jb.*, **LXVIII**, 807–837.

Dobzhansky, T., Wright, S. (1947). Genetics of natural populations X. Dispersal rates in *drosophila pseudoobscura*. *Genetics*, **28**, 304–340.

Fisher, R.A. (1935). The Fiducial Argument in Statistical Inference, *Annals of Eugenics*, **VI**, 91–98.

Ghosh, B.K. (1975). A two-stage procedure for the Behrens-Fisher problem. *Journal of the American Statistical Association*, **70**, 457-462.

Haberle, J.G. (1991). Strength and failure mechanisms of unidirectional carbon fibre-reinforced plastics under axial compression. Unpublished Ph.D. thesis, Imperial College, London, U.K.

Kim, S.H. and Cohen, A.S. (1998). On the Behrens-Fisher problem: a review. *Journal of Educational and Behavioral Statistics*, **23**, 356–377.

Mackowiak, P.A., Wasserman, S.S. and Levine, M.M. (1992). A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl

Reinhold August Wunderlich. *The Journal of the American Medical Association*, **268**, 1578–1580.

Meeusen, W.J., van den Broeck, J. (1977). Efficiency estimation from Cobb Douglas production functions with composed error. *International Economic Review*, **8**, 435–444.

Palmer, A.R. and Strobeck, C. (1986). Fluctuating Asymmetry: Measurement, Analysis, Patterns. *Annual Review of Ecology and Systematics*, **17**, 391–421.

Patil, V.H. (1965). Approximation to the Behrens Fisher distributions. *Biometrika*, **52**, 267–71.

Pewsey, A. (2002). Large-sample inference for the general half-normal distribution. *Communications in Statistics – Theory and Methods*, **31**, 1045–1054.

Pewsey, A. (2004). Improved likelihood based inference for the general half-normal distribution. *Communications in Statistics – Theory and Methods*, **33**, 197–204.

Willink, R. (2004). Approximating the difference of two t-variables for all degrees of freedom using truncated variables. *Australian & New Zealand Journal of Statistics*, **46**, 495–504.

Wiper, M.P., Girón, F.J. and Pewsey, A. (2008). Objective Bayesian inference for the half-normal and half-t distributions. *Communications in Statistics: Theory and Methods*, **37**, 3165–3185.

5. The prior distribution

Objective

Firstly, we study the different methods for elicitation, calibration and combination of *subjective* (expert) prior distributions and secondly, we analyze the different *objective* Bayesian approaches.

Recommended reading

- Berger, J. (2006). The case for objective Bayesian analysis (with discussion). *Bayesian Analysis*, **1**, 385–482.

- Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice (with discussion). *Bayesian Analysis*, **1**, 403–420.

Both articles available from http://ba.stat.cmu.edu/vol01is03.php

Subjective probability distributions and their elicitation

In many real problems, substantive experts with important information are available. In such cases, it is important to be able to convert their information and ideas into probabilities. The techniques for doing this are called *probability elicitation* methods.

Garthwaite et al (2005) observe that there should be four stages in any probability elicitation exercise:

- Setting up: selection and training of experts, investigation of parameters to make judgements about.
- Elicitation of expert opinions: summaries of the experts probability distributions for some parameter.
- Distribution fitting: combining and modeling the expert opinions.
- Feedback: presentation of results to experts to see if they agree or not.

Expert training: problems with the use of expert judgements

Experts often use heuristic methods to make their probability judgements which can induce biases and incoherence.

- motivational biases.
- cognitive biases:
  - ◇ availability
  - ◇ anchoring
  - ◇ representativeness
  - ◇ control

An important part of the training is to try to get the experts to recognize such biases so that they may be eliminated.

Example 22
For each of the following pairs, which causes more deaths per year.

- Stomach cancer or car accidents?
- Tuberculosis or fires?

Cause of death	Choice	Total yearly death rate (USA /1000)	# newspaper articles
cancer	14%	95	46
car accidents	86%	1	137
tuberculosis	23%	4	0
fires	77%	5	0

People have much more information about accidents than about cancer and therefore, this option is more available. See Russo and Shoemaker (1989)

See Tversky and Kahneman (1973) for more examples of the availability bias.

Anchoring and adjustment

Example 23
(Tversky y Kahneman 1974).

The researcher wished to elicit an estimation of the percentage of African countries in the UN. One group of experts were asked the question

Do you think that the percentage is higher or lower than 10%?

and the second group were asked

Do you think that the percentage is higher or lower than 65%?

The members of both groups were then asked to give a point estimate of the percentage.

The mean estimate in group 1 was 25% and the mean in group 2 was 45%. A random and irrelevant anchoring point has influenced the estimates of both groups.

Example 24

Federico is 35 years old, intelligent but not very imaginative and a bit boring. In college, he showed a lot of talent in maths but he wasn't very good at art.

Order the following statements about Federico in terms of their probability (1 = most probable, 8 = least probable).

1. Federico is a doctor and likes to play cards as a hobby.
2. He is an architect.
3. He is an accountant.
4. He plays a jazz instrument.
5. He reads *Marca*.
6. He likes mountaineering.
7. He is an accountant and plays a jazz instrument.
8. He is a journalist.

1. Federico is a doctor and likes to play cards as a hobby.
2. He is an architect.
3. He is an accountant.
4. He plays a jazz instrument.
5. He reads $Marca$.
6. He likes mountaineering.
7. He is an accountant and plays a jazz instrument.
8. He is a journalist.

Most people say that option 3 is the most probable. Moreover, many people say that option 7 is more probable than option 4. This is impossible as for any two events A and B,

$$P(A \cap B) \leq \min\{P(A), P(B)\}.$$

This problem illustrates the representativeness heuristic and also the base rate fallacy. See Kahneman et al (1982) for more examples.

The base rate fallacy

Example 25
(Tversky and Kahneman 1980).

A taxi knocks over a pedestrian in Darlington. In Darlington, only two companies operate taxi services. The first company has green taxis and the second operates blue taxis. Around 85% of the taxis in Darlington are green.

There is a witness to the accident who says that the taxi was blue. When tested under the same climatological conditions as the night of the accident, the witness identifies the two colours correctly in around 80% of the test cases.

What is your estimated probability that the taxi is blue?

The typical response is around 80%. However, if A represents the event that the taxi is blue and a is the event that the witness says it is blue, then from Bayes Theorem:

$$\begin{aligned} P(A|a) &= \frac{P(a|A)P(A)}{P(a|A)P(A) + P(a|\bar{A})P(\bar{A})} \\ &= \frac{0.8 \times 0.15}{0.8 \times 0.15 + 0.2 \times 0.85} = 0.41 \end{aligned}$$

People often ignore the base rate when making their predictions.

Elicitation

Initially we shall consider the simplest problem of eliciting an expert, E's probability, $P_E(A)$, that an event occurs. The simplest approach might appear to be to ask the expert directly for her probability. However, this method does not allow her to think about her probabilities and if she is not statistically trained, will be very hard to implement.

There are two basic alternative approaches based on the use of probability scales or on gambling schemes.

Probability scales

The simplest form of probability scale is a just a straight, unmarked line where the right hand end indicates probability one and the left, probability zero.

The expert is asked to mark a point which represents her probability of a given event.

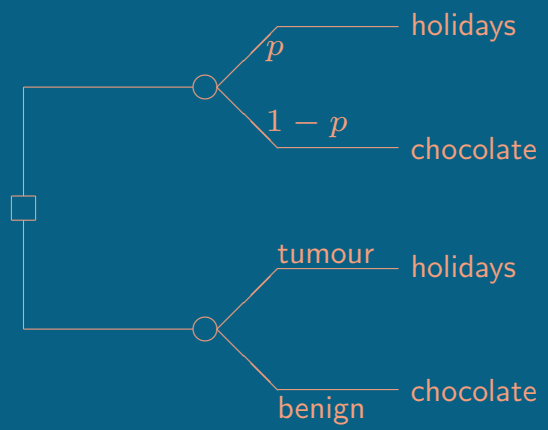
The simple probability scale will not allow the expert to estimate small or large probabilities well. When we are estimating such probabilities, it is better to use an odds scale or a logarithmic scale. Also, it is often useful to include certain guide points on the scale although these might induce *anchoring biases*.

Gambling methods

One typical approach is to use a lottery. Recall, from page 30, that De Finetti (1937) defined subjective probabilities in terms of (*certainty equivalent*) lotteries. Another approach is to consider gambles with one big prize and one negligible prize.

Example 26

Suppose that we wish to elicit a doctor's probability, p_E, that a given patient has a tumour. Then we might ask the doctor to choose to play one of two lotteries as in the following diagram.

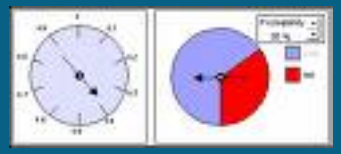


For a given value of p we can check which lottery is preferred and then we can vary p until the doctor is indifferent at some point $p = p_E$. Then $P(\text{tumour}) = p_E$ is the doctor's elicited probability that the patient has the tumour.

There are certain ethical problems with the use of lotteries like this. An alternative is to use *probability wheels*.

Example 27
The illustration shows the *Spinner* probability wheel from *Insight*©.

http://www.stanford.edu/~savage/faculty/savage/InsightInfo.xls



We now ask the doctor to say which event is more likely: that the pointer lands in red or that the patient has a tumour. By varying the size of the coloured sectors we can arrive at a point of indifference when $p = p_E$ is the proportion of the disc coloured red.

Both probability wheel and lottery approaches are basically restricted to assessment of binary probabilities. Also, neither method will work very well if we wish to estimate very small or very large probabilities.

Alternative approaches have been developed based on frequencies (Price 1998) or on attempting to translate verbal expressions such as likely, improbable, etc., into numerical probabilities (Witteman and Renooij 2003). See e.g. Wallsten et al (1993) for a fuller review.

These methods can be extended to elicitation of expert distributions. In these cases, it is most common to elicit *quantiles* from the expert rather than the full distribution function. These quantiles can then be fitted to a given distributional model as required.

Elicitation of (conjugate) prior distributions

There are many possibilities. The worst would be to ask the expert directly about the parameters of the prior distribution.

Example 28

Suppose we are interested in estimating a prior distribution for the probability, p, of heads for a biased coin. In this case, we know that the conjugate prior is beta, $p \sim \mathcal{B}(\alpha, \beta)$, and we wish to derive the values of α and β.

One possibility (Fox 1966) is to ask the expert for a direct estimate of the most likely value of p, i.e. the expert's mode, say p_E, and to state the probability, r_E, that the true value of p lies in an interval $(p_E - Kp_E, p_E + Kp_E)$ where the value of K is fixed by the analyst eliciting the information. Then, assuming that p has a beta prior, $p \sim \mathcal{B}(\alpha, \beta)$, then we can find the values of α and β best representing the expert's judgements by solving

$$
\begin{aligned}
p_E &= \frac{\alpha - 1}{\alpha + \beta - 2} \\
r_E &= \int_{p_E - Kp_E}^{p_E + Kp_E} \frac{1}{B(\alpha, \beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} \, dx
\end{aligned}
$$

In general, it is preferable to ask the expert about *observable* quantities.

Chaloner and Duncan (1983) propose asking the expert to state her mode, x_E, for the number of successes, X, that would occur in a given number, n, of Bernoulli trials and how much less likely it would be that that the number of successes is one less or one more than the mode, say

$$c_E = \frac{P_E(X = x_E - 1)}{P_E(X = x_E)} \quad d_E = \frac{P_E(X = x_E - 1)}{P_E(X = x_E)}.$$

Then, recalling that the marginal distribution of X supposing a beta prior is beta-binomial, the parameters may be estimated by solving the following system of equations for α and β.

$$c_E = \frac{(n - x_E)(x_E + \alpha)}{(x_E + 1)(n - x_E + \beta - 1)} \quad \text{and} \quad d_E = \frac{x_E(n - x_E + \beta)}{(n - x_E + 1)(n - \alpha + 1)}.$$

Many other approaches have been considered. See e.g. Hughes and Madden (2002).

Evaluating the quality of expert forecasts

Typically used criteria (Lichtenstein et al 1982) are the following:

- *Honesty*: We want the expert to tell the truth.
- *Coherence*: Her forecasts should satisfy the laws of probability.
- *Consistency*: If she doesn't receive new information, then her predictions shouldn't change.
- *Calibration*: It should rain on around 50% of the days when the expert says $P(\text{rain}) = 0.5$.
- *Informativeness:* If, in Madrid, it rains on around 50 days per year, an expert who says

  $$P(\text{rain tomorrow}) = 50/364$$

  every day isn't very informative.

Honesty and strictly proper scoring rules

Suppose that we wish to elicit the expert's true probability, p_E, that some event A occurs. One method of encouraging the expert to be honest is to pay her a quantity $R(A, p)$ which depends upon the occurrence or not of A and the expert's stated probability, p.

How should we define $R(A, p)$?.

We suppose that the expert wishes to maximize her expected income. If p_E is her true probability, her expected income if she states a probability p is

$$p_E R(1, p) + (1 - p_E) R(0, p).$$

Definition 9

A *(strictly) proper scoring rule* (Savage 1971) is a scoring rule $R(A, p)$ whereby the expert maximizes his expected income if (and only if) $p = p_E$.

Example 29

Suppose that $R(A,p) = 1 - |A - p|$. Then, the expert's expected earnings if she states a probability p are

$$\begin{aligned} E[R] &= p_E\,(1 - |1 - p|) + (1 - p_E)\,(1 - |0 - p|) \\ &= p_E p + (1 - p_E)(1 - p) \\ &= 1 - p_E + (2p_E - 1)p \end{aligned}$$

Therefore $R(A,p)$ is not a proper scoring rule as the expert maximizes her expected earnings by stating $p = 1$ (0) if $p_E > (<) 0.5$.

The Brier score

Example 30
$R(A, p) = 1 - (A - p)^2$ is the Brier (1950) score.

$$\begin{aligned} E[R] &= p_E\left(1-(1-p)^2\right) + (1-p_E)\left(1-p^2\right) \\ &= 1 - p_E + 2pp_E - p^2 \\ &= 1 - p_E + p_E^2 - (p - p_E)^2 \end{aligned}$$

which is maximized by setting $p = p_E$. Therefore, R is a strictly proper scoring rule.

There are many other proper scoring rules, see e.g. Winkler (1986) and proper scoring rules have also been developed for continuous variables and quantiles. See e.g. Buehler (1971) and Matheson and Winkler (1976).

Example 31

Suppose that E is asked to state a point estimator, say e, for a variable X. Then consider the scoring rule

$$R(X,e) = \begin{cases} a(e-x) & \text{if } e < x \\ b(x-e) & \text{if } e > x \end{cases}$$

Let $p_E(x)$ be the expert's true distribution for X:

$$\begin{aligned} E[R(X,e)] &= \int R(x,e)p_E(x)\,dx \\ &= a\int_e^\infty (e-x)p_E(x)\,dx + b\int_{-\infty}^e (x-e)p_E(x)\,dx \\ &= ae(1-F_E(e)) - a\int_e^\infty xp_E(x)\,dx + \\ &\quad b\int_{-\infty}^e xp_E(x)\,dx - beF_E(e) \end{aligned}$$

$$
\begin{aligned}
\frac{dE[R(X,e)]}{de} &= a(1-F_E(e)) - aep_E(e) + aep_E(e) - \\
&\quad bep_E(e) - bF_E(e) + bep_E(e) \\
0 &= a(1-F_E(\hat{e})) - bF_E(\hat{e}) \\
F_E(\hat{e}) &= \frac{a}{a+b}
\end{aligned}
$$

The expert maximizes her experted gains if she states her $b/(a+b) \times 100\%$ quantile. See Raiffa and Schlaifer (1961).

The use of proper scoring rules to encourage honesty seems somewhat artificial. However, they may also be used *a posteriori* as evaluation tools for expert probabilities.

Numerical measures of expert quality

Suppose that the expert supplies probabilities $\mathbf{p}_E$ for a sequence of Bernoulli events $X_1, \ldots, X_n$. Given the data $\mathbf{x}$, we wish to evaluate the quality of her predictions.

Consider the Brier score

$$R(X, p_E) = 1 - (X - p_E)^2.$$

Then, given the data, we can calculate the statistic

$$R(\mathbf{x}, \mathbf{p}_E) = \frac{1}{n} \sum_{i=1}^{n} R(x_i, p_{Ei}) = 1 - \sum_{i=1}^{n} (x_i - p_{Ei})^2$$

which is a measure of the average quality of her predictions.

This measure can be divided into a measure of calibration and a measure of information.

Following Murphy (1973), assume that the expert uses the probability p_j a total of n_j times, with a frequency of f_j successes and a relative frequency of $r_j = f_j/n_j$ successes for $j = 1, \ldots k$. Then:

$$\begin{aligned} R(\mathbf{x}, \mathbf{p}) &= 1 - \frac{1}{n} \sum_{j=1}^{k} \left(f_j (1 - p_j)^2 + (n_j - f_j)(0 - p_j)^2 \right) \\ &= 1 - \frac{1}{n} \sum_{j=1}^{k} n_j \left(r_j (1 - p_j)^2 + (1 - r_j)(0 - p_j)^2 \right) \end{aligned}$$

Now we can prove the following theorem.

Theorem 25
$R(\mathbf{x}, \mathbf{p}) = 1 - C(\mathbf{x}, \mathbf{p}) - I(\mathbf{x}, \mathbf{p})$ where

$$C(\mathbf{x}, \mathbf{p}) = \frac{1}{n} \sum_{j=1}^{k} n_j (r_j - p_j)^2 \quad \text{is a measure of calibration}$$

$$I(\mathbf{x}, \mathbf{p}) = \frac{1}{n} \sum_{j=1}^{k} n_j r_j (1 - r_j) \quad \text{is a measure of information}$$

Proof

$$
\begin{aligned}
R(\mathbf{x},\mathbf{p}) &= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j(1-p_j)^2+(1-r_j)(0-p_j)^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j-2r_jp_j+p_j^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j-r_j^2+r_j^2-2r_jp_j+p_j^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j(1-r_j)+(r_j-p_j)^2\right)\\
&= 1-I(\mathbf{x},\mathbf{p})-C(\mathbf{x},\mathbf{p})
\end{aligned}
$$

C has the following properties:

- $0 \leq C \leq 1$
- $C = 0$ if and only if $r_j = p_j$ for $j = 1, \ldots, k$.
- For a well calibrated expert, when $n \to \infty$, $C \to 0$.
- C is large if the observed relative frequencies r_i are very different from the expert's stated probabilities p_i.

I has the following properties:

- $0 \leq I \leq 0.25$.
- $I = 0$ if, for all p_j, the relative frequency $r_j = 0$ or 1.
- $I = 0.25$ if, for all p_j, then $r_j = 0.5$.

Any strictly proper scoring rule can be divided up into calibration and information measures in a similar way. See e.g. De Groot and Fienberg (1982).

Example 32

Wiper (1987,1990) gave 12 experts a set of 50 statements to study. The experts were asked to state whether each statement was true or false and to give their probabilities that they were correct as a percentage between 50% (= no idea) and 99%.

We shall assume here that the experts' stated probabilities for all events belong to the class $\mathbf{p} = \{0.53, 0.64, 0.75, 0.86, 0.97\}$. Then the following table gives the related absolute and relative frequencies.

E	p_i	0.53	0.64	0.75	0.86	0.97
2	n_i	25	6	6	5	8
	f_i	15	4	4	3	8
	r_i	0.6	0.67	0.67	0.6	1.0
3	n_i	25	5	10	5	5
	f_i	16	1	3	2	4
	r_i	0.64	0.2	0.3	0.4	0.8
10	n_i	10	5	15	1	19
	f_i	6	2	5	0	15
	r_i	0.6	0.4	0.3	1.0	0.79

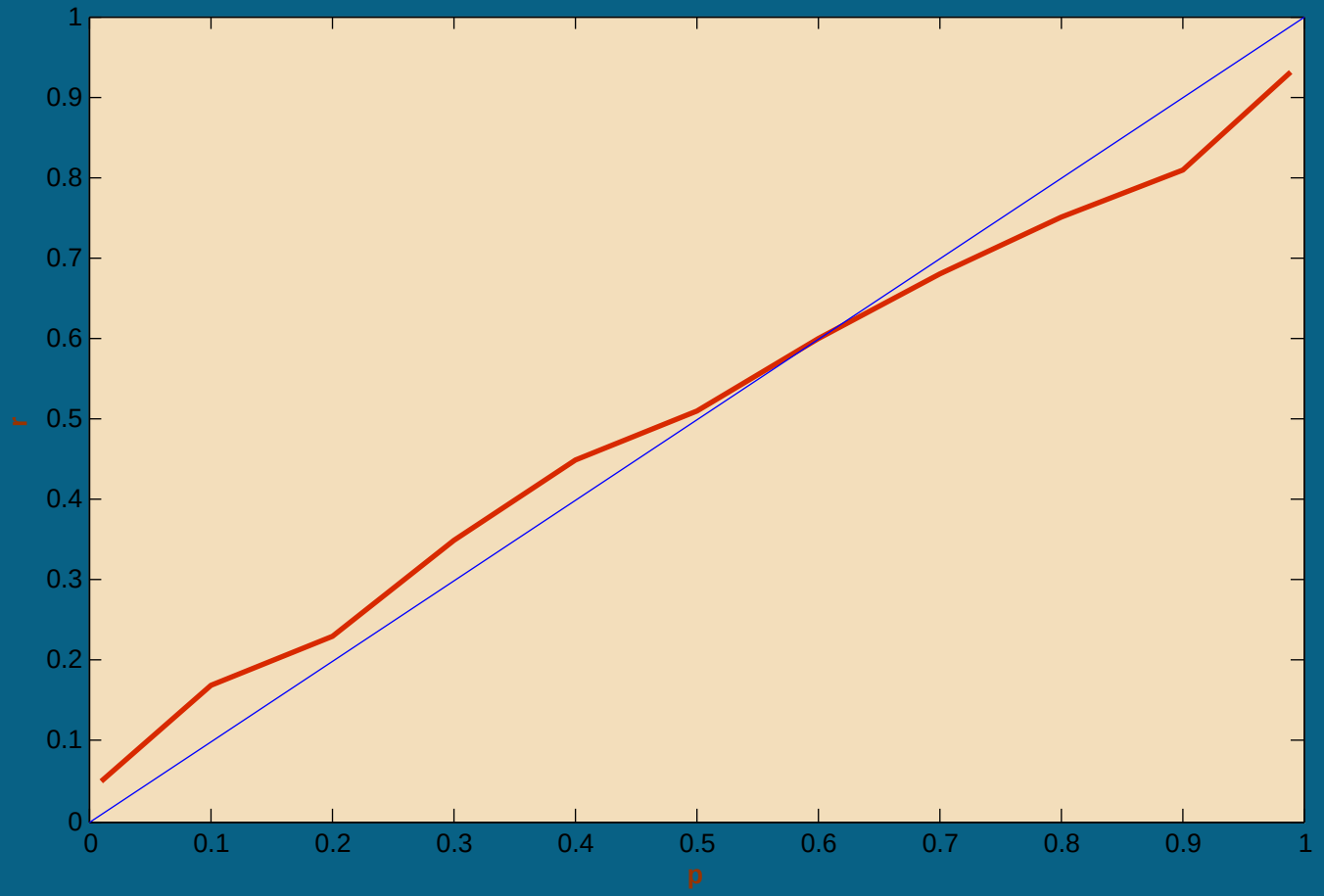
The following table shows the calibration, information and Brier scores for each expert.

E	C	I	Brier
2	.0093	.1973	.7934
3	.0900	.2132	.6968
10	.1059	.2018	.6923

We can see that expert 2 is better calibrated but less informative than the other experts.

A visual manner of illustrating expert calibration is provided by the calibration curve. This is simply a graph of observed frequencies, r_j, against the different probabilities used, p_j.

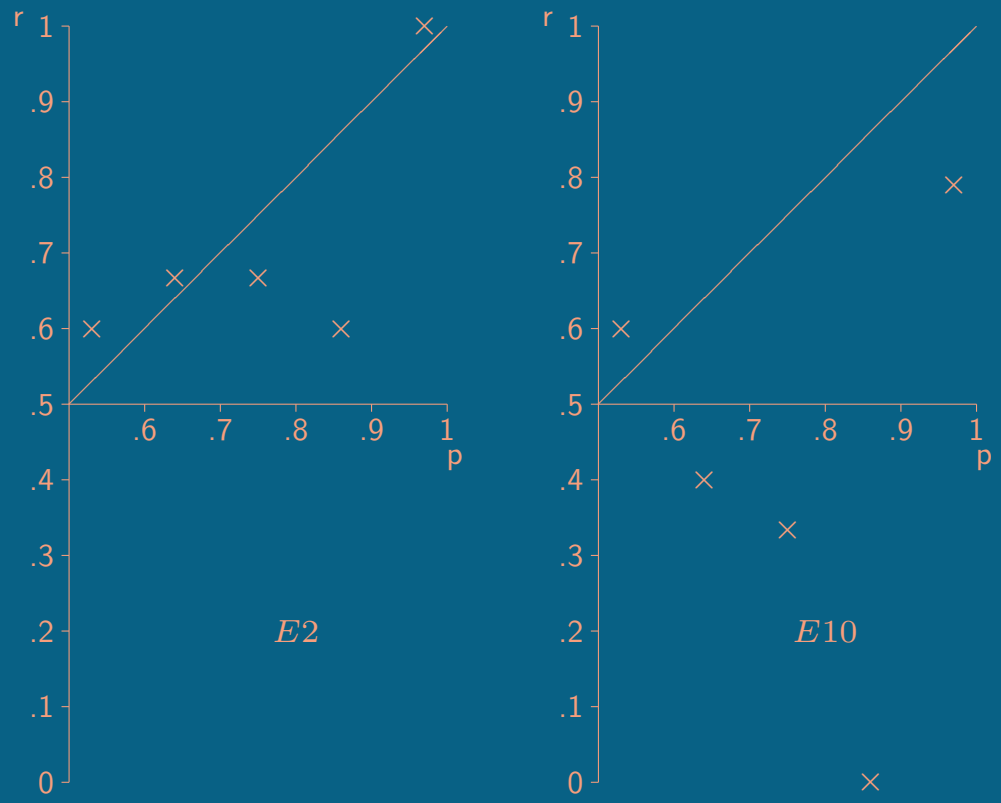
The calibration curve



For a well calibrated expert, the curve approximates the 45 degree line.

Example 33

Returning to Example 32, the calibration curves of experts 2 and 10 are given in the following figure.



Cooke's approach

Cooke et al (1988) and Cooke (1991) have developed alternative measures of calibration and information based on classical p-values which can be applied to predictions for both discrete and continuous variables.

Suppose that an expert uses the probability p_j a total of n_j times for $j = 1, 2, \ldots, k$. Theoretically, if the expert is well calibrated, about the total frequency of events $\{X = 1\}$ that occur should be around $n_j \times p_j$.

A χ^2 test can be set up to test whether the observed relative frequencies $r_1, \ldots, r_k$ could be generated from the theoretical distribution $p_1, \ldots, p_k$.

Thus, to test $H_0 : r \sim p$ against the alternative $H_1 : r \nsim p$ we calculate the chi-squared statistic

$$S = \sum_{j=1}^{k} n_j \frac{(r_j - p_j)^2}{r_j}.$$

The hypothesis that the expert is well calibrated can then be accepted or rejected by comparing S with tables of the χ^2_k distribution.

Example 34
The following table shows the p-values for each expert in Example 32.

E	2	3	10
p	.7	.00	.00

It seems that only expert 2 is reasonably well calibrated.

Cooke (1991) derives a theory of scoring rules based on combining the p-value with a measure of information. When the expert makes forecasts for continuous variables, then an alternative is a Kolmogorov Smirnov test. See Wiper et al (1994).

Objective Bayesian methods

Sometimes we wish to use a prior distribution which does not include (much) subjective information, because

- we don't know anything about the problem at hand,
- we would like to be *objective*.

In such situations, we should choose a *non-informative* prior distribution.

However, there are many possible elections. Which is the most useful?

Uniform priors

Bayes (1763) and Laplace (1812) generally employed uniform prior distributions, as justified by the principle of insufficient reason, see page 23. This is fine in finite dimensional problems, but if the parameter space Θ is continuous or uncountable, then such prior distributions are $improper$. In practice, this is not a serious problem, as long as the posterior distribution as calculated via Bayes theorem can be shown to exist.

An important problem with the general use of uniform priors however is lack of invariance to transformation.

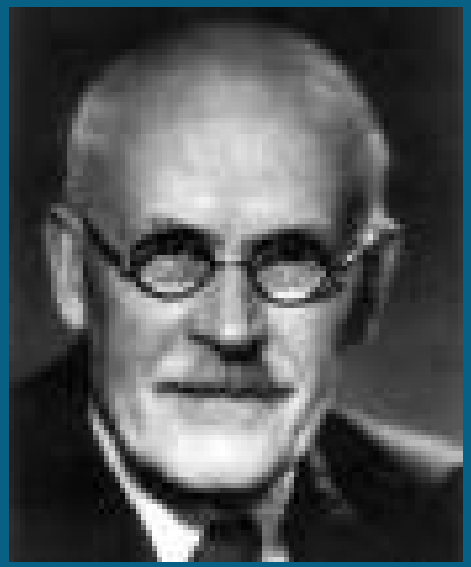
Example 35

Suppose that we set $p(\theta) \propto 1$ and define the transformed variable $\phi = \frac{1}{\theta}$. Then the implied prior distribution for ϕ is

$$p(\phi) \propto \frac{1}{\phi^2}$$

which is not uniform. Thus, the use of the uniform distribution to represent lack of knowledge is inconsistent. If we know nothing about θ, then we should know nothing about ϕ.

Jeffreys priors



Jeffreys

Jeffreys (1946) introduced a prior distribution which possesses an invariance property.

Definition 10
Let $X|\theta \sim f(\cdot|\theta)$ where θ is one dimensional. The Jeffreys prior for θ is

$$p(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta) = -E_X\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right]$ is the expected Fisher information.

The following theorem shows that if the Jeffreys prior for θ is used, then the implied prior for the transformed parameter $\phi = \phi(\theta)$ is the Jeffreys prior for ϕ.

Theorem 26
If $\phi = \phi(\theta)$, and $p(\theta)$ is the Jeffreys prior for θ, then the implied prior for ϕ is the Jeffreys prior

$$p(\phi) \propto \sqrt{I(\phi)}$$

where $I(\phi) = -E_X\left[\frac{d^2}{d\phi^2}\log f(X|\phi)\right]$ is the expected Fisher information when the distribution of X is reparameterized in terms of ϕ.

Proof Firstly, we shall prove that $E\left[\frac{d}{d\theta}\log f(X|\theta)\right] = 0$.

$$\begin{aligned} E_X\left[\frac{d}{d\theta}\log f(X|\theta)\right] &= \int\left[\frac{d}{d\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\ &= \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx \\ &= \int f'(x|\theta)\,dx \\ &= \frac{d}{d\theta}\int f(x|\theta)\,dx = 0. \end{aligned}$$

Secondly we shall prove that

$$E\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] = -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right].$$

We have

$$
\begin{aligned}
E_X\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] &= \int \frac{d^2}{d\theta^2}\log f(x|\theta)\,dx \\
&= \int \frac{d}{d\theta}\left[\frac{d}{d\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\
&= \int \frac{d}{d\theta}\frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx \\
&= \int \left[\frac{f''(x|\theta)}{f(x|\theta)} - \left(\frac{f'(x|\theta)}{f(x|\theta)}\right)^2\right] f(x|\theta)\,dx \\
&= \int \frac{d^2}{d\theta^2} f(x|\theta)\,dx - \int \left(\frac{d}{d\theta}\log f(x|\theta)\right)^2 f(x|\theta)\,dx \\
&= -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]
\end{aligned}
$$

Now finally, we have that

$$
\begin{aligned}
p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \quad \text{and squaring,} \\
p(\phi)^2 &= p(\theta)^2\left|\frac{d\theta}{d\phi}\right|^2 \\
&\propto -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]\left|\frac{d\theta}{d\phi}\right|^2 \\
&\propto E\left[\left(\frac{d}{d\phi}\log f(X|\phi)\right)^2\right] \\
p(\phi) &\propto \sqrt{I(\phi)}.
\end{aligned}
$$

Example 36
$X|\theta \sim \mathcal{BI}(n, \theta)$.

$$\begin{aligned}
\log f(X|\theta) &= c + X \log \theta + \\
&\quad +(n - X) \log(1 - \theta) \\
\frac{d}{d\theta} \log f(X|\theta) &= \frac{X}{\theta} - \frac{(n - X)}{(1 - \theta)} \\
\frac{d^2}{d\theta^2} \log f(X|\theta) &= -\frac{X}{\theta^2} - \frac{(n - X)}{(1 - \theta)^2} \\
E\left[\frac{d^2}{d\theta^2} \log f(X|\theta)\right] &= -n\left(\frac{1}{\theta} + \frac{1}{1 - \theta)}\right) \\
I''(\theta) &\propto \frac{1}{\theta(1 - \theta)}
\end{aligned}$$

Therefore, the Jeffreys prior is $p(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}}$, that is $\theta \sim \mathcal{B}(1/2, 1/2)$. This is a *proper* prior, unlike Haldane's prior which we saw earlier.

Example 37

$X|\mu \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$, with ϕ known.

$$\begin{aligned} \log f(X|\mu) &= c - \frac{\phi}{2}(X-\mu)^2 \\ \frac{d}{d\mu}\log f(X|\mu) &= \phi X - \phi\mu \\ \frac{d^2}{d\mu^2}\log f(X|\mu) &= -\phi \end{aligned}$$

We have $p(\mu) \propto 1$, a uniform distribution.

Example 38

Suppose now that $X|\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$ where μ is known.

$$\begin{aligned}
\log f(X|\phi) &\propto \frac{1}{2}\log\phi - \phi\frac{(X-\mu)^2}{2} \\
\frac{d}{d\phi}\log f(X|\phi) &= \frac{1}{2\phi} + \frac{(X-\mu)^2}{2} \\
\frac{d^2}{d\phi^2}\log f(X|\phi) &= -\frac{1}{2\phi^2}
\end{aligned}$$

and the Jeffreys prior for ϕ is

$$p(\phi) \propto \frac{1}{\phi}.$$

If σ is the standard deviation, then $\phi = \frac{1}{\sigma^2}$ and $\frac{d\phi}{d\sigma} = -\frac{2}{\sigma^3}$ so the Jeffreys prior for σ is

$$p(\sigma) \propto \sigma^2\frac{2}{\sigma^3} \propto \frac{1}{\sigma}.$$

Jeffreys priors in multivariate problems

It is possible to extend the definition of a Jeffreys prior to the case when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is multivariate, by defining $p(\boldsymbol{\theta}) \propto \sqrt{I(\boldsymbol{\theta})}$ as earlier, where the expected Fisher information is now given by

$$I(\boldsymbol{\theta}) = |E_X [\mathbf{J}(\boldsymbol{\theta})]| \quad \text{where}$$

$$\mathbf{J}_{ij} = \frac{d^2}{d\theta_i d\theta_j} \log f(X|\boldsymbol{\theta}).$$

In some cases, the multivariate Jeffreys prior seems reasonable.

Example 39
Let $X|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ have a (k dimensional) multinomial distribution. Then

$$\begin{aligned}
\log f(X|\boldsymbol{\theta}) &= c + \sum_{i=1}^{k} x_i \log \theta_i \\
\frac{d}{d\theta_i} \log f(X|\boldsymbol{\theta}) &= \frac{X_i}{\theta_i} \\
\frac{d^2}{d\theta_i^2} \log f(X|\boldsymbol{\theta}) &= -\frac{X_i}{\theta_i^2} \\
\frac{d^2}{d\theta_i d\theta_j} \log f(X|\boldsymbol{\theta}) &= 0 \quad \text{for } i \neq j. \\
E\left[\frac{d^2}{d\theta_i^2} \log f(X|\boldsymbol{\theta})\right] &= -\frac{m}{\theta_i} \\
I(\boldsymbol{\theta}) &= \frac{m^k}{\prod_{i=1}^{k} \theta_i}
\end{aligned}$$

and the Jeffreys prior is $p(\boldsymbol{\theta}) \propto \prod_{i=1}^{k} \frac{1}{\sqrt{\theta_i}}$ which is Dirichlet, $\boldsymbol{\theta} \sim \mathcal{D}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$.

In many more cases, the multivariate Jeffreys prior is less natural.

Example 40

Let $X|\mu,\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$. Then, from Examples 37 and 38,

$$\begin{aligned}
\frac{d^2}{d\mu^2}\log f(X|\mu,\phi) &= -\phi \\
\frac{d^2}{d\phi^2}\log f(X|\mu,\phi) &= -\frac{1}{2\phi^2} \quad \text{and also} \\
\frac{d^2}{d\mu d\phi}\log f(X|\mu,\phi) &= -(X-\mu) \\
E[\mathbf{J}] &= -\begin{pmatrix} \phi & 0 \\ 0 & \frac{1}{2\phi^2} \end{pmatrix}
\end{aligned}$$

and therefore, the Jeffreys prior is $p(\mu,\phi) \propto \frac{1}{\sqrt{\phi}}$ which is not the prior we have used earlier.

Maximum entropy priors

Jaynes

The idea (Jaynes 1968,1983) is to find the least informative prior distribution in the presence of partial information.

Entropy

Assume that θ is univariate and discrete. If $p(\theta)$ is any distribution for θ, then we can define

$$e(p) = -\sum_{i \in \Theta} p(\theta_i) \log p(\theta_i)$$

to be the *entropy* of the distribution.

If $p(\theta = \theta_i) = 1$ for some value $\theta_i \in \Theta$, then, $e(p) = 0$ and there is zero uncertainty or minimum entropy. On the contrary, if $p(\theta_i) = 1/|\Theta|$, i.e. a uniform distribution, then

$$e(p) = -\sum_{i \in \Theta} \frac{1}{|\Theta|} \log \frac{1}{|\Theta|} = \log |\Theta|,$$

the maximum entropy.

Maximum entropy (maxent) distributions are minimum information distributions.

In many practical situations, we may only wish to fix certain characteristics of the prior distribution, e.g. quantiles or moments and apart from this, let the prior be as uninformative as possible.

Suppose that we have partial information about θ in the form

$$E[g_k(\theta)] = \sum_{i \in \Theta} p(\theta_i) g_k(\theta_i) = \mu_k$$

for $k = 1, \ldots, m$. This includes fixed moments, e.g. $g_1(\theta) = \theta$ and quantiles

$$g_k(\theta) = I_{(-\infty, z_k]} \Rightarrow E[g_k(\theta)] = p(\theta \leq z_k).$$

Given these restrictions, the following theorem provides the form of the maximum entropy prior.

Theorem 27
Given the partial information

$$E[g_k(\theta)] = \sum_{i \in \Theta} p(\theta_i) g_k(\theta_i) = \mu_k$$

for $k = 1, \ldots, m$, then the maxent prior is

$$p(\theta_i) = \frac{\exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}{\sum_{j \in \Theta} \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_j)\right)}$$

where the constants λ_k can be determined from the information.

Proof See Jaynes (1968). ∎

Example 41

Let $X|N \sim \mathcal{BI}(N, 1/2)$. Suppose that we know that $N \geq 1$ and that we fix the mean to be $E[N] = 10$.

We shall try to calculate the maxent distribution for N.

$$\begin{aligned} P(N = n) &= \frac{\exp(\lambda_1 n)}{\sum_{j=1}^{\infty} \exp(\lambda_1 j)} \\ &= \exp(\lambda_1 n) \frac{1 - \exp(\lambda_1)}{\exp(\lambda_1)} \\ &= (1 - e^{\lambda_1}) \exp(\lambda_1 (n-1)) \end{aligned}$$

Thus, $N - 1$ has a geometric density with parameter $1 - e^{\lambda_1}$ and therefore

$$E[N] = 1 + \frac{e^{\lambda_1}}{1 - e^{\lambda_1}}$$

and fixing $E[N] = 10$, we have $e^{\lambda_1} = \frac{9}{10}$, i.e. the maxent prior for N is $N - 1 \sim \mathcal{GE}(9/10)$.

Maxent priors for continuous variables

The extension to continuous variables is more complicated because the definition of entropy

$$e(p) = -\int p \log p \, d\mu$$

depends on the base measure μ.

One possibility (Jaynes 1968) is to define

$$e(p) = -\int p(\theta) \log \frac{p(\theta)}{p_0(\theta)} \, d\theta$$

where $p_0(\theta)$ is the Jeffreys prior for θ. Then, given the restrictions $E[g_k(\theta)] = \lambda_k$, the maxent prior is

$$p(\theta) = \frac{p_0(\theta) \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right)}{\int p_0(\theta) \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right) \, d\theta}$$

analogous to the discrete case.

Unfortunately, in some cases, it is possible that no maxent prior distribution exists.

Example 42
Let $X|\mu \sim \mathcal{N}(\mu, 1)$ and suppose that we fix the prior mean to be $E[\mu] = m$. We have seen in Chapter 4 that the Jeffreys prior is $p(\mu) \propto 1$. Therefore, the maxent distribution is

$$p(\mu) = \frac{\exp(\lambda_1 \mu)}{\int_{-\infty}^{\infty} \exp(\lambda_1 \mu) d\mu}$$

and there is no solution to this integral.

Reference priors

Bernardo

This, the most general approach, was developed by Bernardo (1979). It is based on maximizing the expected information about $\boldsymbol{\theta}$ to be provided by a given experiment.

Consider first the case when θ is one dimensional.

Then, given the prior distribution $p(\theta)$, the expected information about θ to be gained by observing a sample $\mathbf{X}$ of size n, where $X|\theta \sim f(\cdot|\theta)$ is defined by

$$I(p(\theta)) = \int f(\mathbf{x}) \int p(\theta|\mathbf{x}) \log \frac{p(\theta|\mathbf{x})}{p(\theta)} \, d\theta \, d\mathbf{x}$$

where $f(\mathbf{x}) = \int f(\mathbf{x}|\theta) p(\theta) \, d\theta$ and $p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{f(\mathbf{x})}$.

Then, the *reference prior* is defined to be the prior $p(\theta)$, within the class of admissible priors, which maximizes the asymptotic limit of the expected information $I(p(\theta))$ as the sample size n goes to infinity.

It can be shown that the maximum entropy and Jeffreys priors correspond to particular cases of reference priors.

Reference priors in multivariate problems

Suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is a nuisance parameter. Then the reference prior approach has two steps:

1. Calculate the reference prior $p(\theta_2|\theta_1)$ as above.

2. If this is proper, then θ_2 can be integrated out of the density of X and the reference prior of θ_1 can be found as above. If not, then the procedure can be performed in a limiting way. See Bernardo (1979).

Other non-informative prior distributions

There are a number of other approaches to defining non-informative priors:

- Limiting forms of conjugate priors.

  This is the method we used in chapters 2 and 3.

- Priors based on the data translated likelihood. Box and Tiao (1973).

  This approach shows how to define the transformation $\phi = \phi(\theta)$ so that a uniform prior for ϕ is justified. The resultant distributions are Jeffreys priors.

- Methods based on symmetry, e.g. Haar priors.

- Others: see Yang and Berger (1997) and Kass and Wassermann (1996).

Problems with the use of non-informative prior distributions

There are various theoretical and practical difficulties with the use of non-informative priors. Firstly, when improper prior distributions are used, it is important to show that the posterior distribution is proper.

Example 43
Let $X|\theta \sim \mathcal{BI}(n, \theta)$ and suppose we use Haldane's prior $p(\theta) \propto \frac{1}{\theta(1-\theta)}$. Then, if we observe $X = 0$ or $X = n$, then the posterior distribution is improper.

This is particularly important in modern Bayesian models where high dimensional integration is often carried out using Gibbs sampling or MCMC.

Example 44
Consider the simple random effects model

$$\begin{aligned} y_{ij} &= \beta + \mu_i + \epsilon_{ij} \quad \text{where} \\ \epsilon_{ij}|\phi_\epsilon &\sim \mathcal{N}\left(0, \frac{1}{\phi_\epsilon}\right) \end{aligned}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ where $\sum_{i=1}^{k} n_i = n$, and suppose that we use the improper priors

$$\begin{aligned} p(\beta) &\propto 1, \qquad \boldsymbol{\mu} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi_\mu}\mathbf{I}\right) \\ p(\phi_\epsilon) &\propto \frac{1}{\phi_\epsilon}, \qquad p(\phi_\mu) \propto \frac{1}{\phi_\mu}. \end{aligned}$$

Then, given the sample data, $\mathbf{y}$, we can show that the conditional posterior distributions are

$$\beta|\mathbf{y},\boldsymbol{\mu},\phi_\epsilon,\phi_\mu \sim \mathcal{N}\left(\bar{y}-\frac{1}{n}\sum_{i=1}^{k}n_i\mu_i,\frac{1}{n\phi_\epsilon}\right)$$

$$\boldsymbol{\mu}|\mathbf{y},\beta,\phi_\epsilon,\phi_\mu \sim \mathcal{N}\left(\begin{pmatrix}\frac{n_1\phi_\epsilon(\bar{y}_1-\beta)}{n_1\phi_\epsilon+\phi_\mu}\\ \vdots \\ \frac{n_k\phi_\epsilon(\bar{y}_k-\beta)}{n_k\phi_\epsilon+\phi_\mu}\end{pmatrix},\begin{pmatrix}\frac{1}{n_1\phi_\epsilon+\phi_\mu} & 0 & \dots & 0\\ \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \frac{1}{n_k\phi_\epsilon+\phi_\mu}\end{pmatrix}\right)$$

$$\phi_\epsilon|\mathbf{y},\beta,\boldsymbol{\mu},\phi_\mu \sim \mathcal{G}\left(\frac{n}{2},\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\beta-\mu_i)^2}{2}\right)$$

$$\phi_\mu|\mathbf{y},\beta,\boldsymbol{\mu},\phi_\epsilon \sim \mathcal{G}\left(\frac{k}{2},\frac{\sum_{i=1}^{k}\mu_i^2}{2}\right)$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}$ and $\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$.

Now, a Gibbs sampler could be set up by sampling sequentially from the various conditional distributions. However, the results do not make sense as the joint posterior distribution in this case can be shown to be improper. See e.g. Hill (1965).

There are a number of published papers using MCMC methods where the results are, in reality, meaningless as the posterior distributions are really improper.

The likelihood principle

The use of Jeffreys priors does not satisfy the likelihood principle.

Example 45

Suppose that we are going to generate binomial data $X|\theta \sim \mathcal{BI}(n, \theta)$. Then, from Example 36 we know that the Jeffreys prior is $\theta \sim \mathcal{B}(1/2, 1/2)$. Now suppose that we change the experimental design so that we will now generate negative binomial data. Therefore:

$$\begin{aligned}
\log f(X|\theta) &= c + r \log \theta + X \log(1-\theta) \\
\frac{\partial \log f(X|\theta)}{\partial \theta} &= \frac{r}{\theta} - \frac{X}{1-\theta} \\
\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} &= -\frac{r}{\theta^2} - \frac{X}{(1-\theta)^2} \\
-E\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right] &= \frac{r}{\theta^2} + \frac{r}{\theta(1-\theta)} = \frac{r}{\theta^2(1-\theta)}
\end{aligned}$$

The Jeffreys prior is

$$p(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}}.$$

Thus, Jeffreys prior depends on the experimental design and, if we observe 9 heads in 12 tosses of the coin, we need to know the design before the posterior distribution can be calculated. The posterior is $\theta|\mathbf{x} \sim \mathcal{B}(9.5, 3.5)$ given binomial data and $\mathcal{B}(9, 3.5)$ given negative binomial data.

Other problems

- Inadmissibility: Bayesian inference (with proper priors) leads to admissible estimators, but the use of improper priors can lead to inadmissible estimators.

  Example 46
  Suppose that $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}\right)$ and that we use a uniform prior $p(\boldsymbol{\theta}) \propto 1$. Then given an observation, $\mathbf{x}$, the posterior is $\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I})$ and the posterior mean, $\mathbf{x}$ is an inadmissable estimate of $\boldsymbol{\theta}$ if the dimension of $\mathbf{X}$ is greater than 2.

- Marginalization paradoxes (Dawid et al 1973), strong inconsistency and incoherence (Stone and Dawid 1972, Stone 1982).

References

Bayes, T.R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370–418.

Berger, J. (2006). The case for objective Bayesian analysis (with discussion). *Bayesian Analysis*, **1**, 385482.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Buehler, R. (1971). Measuring information and uncertainty. In Godambe, V. and Sprott, D. eds. *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B*, **41**, 113–147.

Box, G.E. and Tiao, G.C. (1973). *Bayesian inference and statistical analysis*. Reading, MA: Addison-Wesley.

Chaloner, K. and Duncan, G.T. (1983). Assessment of a beta distribution: PM elicitation. *The Statistician*, **32**, 174-180.

Cooke, R.M. (1991). *Experts in Uncertainty*. New York: Oxford University Press.

Cooke, R.M., Mendel, M. and Thijs, W. (1988). Calibration and information in expert resolution: a classical approach. *Automatica*, **24**, 87–94.

Dawid, P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **35**, 189–233.

DeGroot, M. and Fienberg, S. (1982). Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics, III*, eds. Gupta, S. and Berger, J. New York: Academic Press.

Finetti, R. de (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de lInstitut Henri Poincaré*, **7**, 1-68.

Fox, B.L. (1966). A Bayesian approach to reliability assessment. Memorandum **RM-5084-NASA**, The Rand Corporation, Santa Monica, USA.

Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–701.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice (with discussion). *Bayesian Analysis*, **1**, 403420.

Hill, J.M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, **60**, 806–825.

Hughes, G. and Madden, L.V. (2002). Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence. *Crop Protection*, **21**, 203-215.

Kahneman, D. Slovic, P. and Tversky, A. (1982). *Judgment under uncertainty: heuristics*

and biases. Cambridge: University Press.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. on Systems Science and Cybernetics*, **4**, 227–241.

Jaynes, E.T. (1983). *Papers on Probability, Statistics and Statistical Physics*. Ed. R.D. Rosenkrantz. Dordrecht: Reidel.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, **A. 186**, 453-.461.

Laplace, P.S. (1812). Teoría analítica de las probabilidades.

Lichtenstein, S., Fischoff, B. and Phillips, L. (1982). Calibration of probabilities: the state of the art to 1980. In Kahneman, D., Slovic, P. and Tversky, A., eds. *Judgement under uncertainty: heuristics and biases*. Cambridge: University Press.

Kass, R.E. and Wassermann (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Matheson, J. and Winkler, R.L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

Murphy, A. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595–600.

Price, P.C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes*, **76**, 277-297.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University Press.

Russo, J.E. and Shoemaker, P.J.H. (1989). *Decision traps*. New York: Simon and Schuster.

Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 781–801.

Stone, M. (1982), Review and analysis of some inconsistencies related to improper priors and finite additivity. *Logic, Methodology, and Philosophy of Science VI: Proceedings of the Sixth International Congress*, Hannover 1979, 413–426, Amsterdam: North-Holland.

Stone, M. and Dawid, A.P. (1972), Un-Bayesian implications of improper Bayes inference in routine statistical problems, *Biometrika*, **59**, 369–375.

Tversky, A. and Kahneman, D. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–51.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1130.

Tversky, A. and Kahneman, D. (1980). Causal Schemas in Judgment Under Uncertainty. In Fischbein, M. ed. *Progress in Social Psychology*, Hillsdale, NJ: Lawrence Erlbaum, 49–72.

Wallsten, T.S., Budescu, D.V., and Zwick, R. (1993). Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science*, **39**,

176–190.

Winkler, R.L. (1986). On good probability appraisers. In Goel, P. and Zellner, A. eds. *Bayesian Inference and Decision Techniques*. New York: Elsevier.

Wiper, M.P. (1987). *The expert problem*. M.Sc. Dissertation. Department of Statistics, Manchester University.

Wiper, M.P. (1990). *Calibration and use of expert probability judgements*. Ph.D. Thesis. Department of Computer Studies, Leeds University.

Wiper, M.P., French, S. and Cooke, R. (1994). Hypothesis based calibration scores. *The Statistician*, **43**, 231–236.

Witteman, C. and Renooij, S. (2003). Evaluation of a verbalnumerical probability scale. *International Journal of Approximate Reasoning*, **33**, 117-131.

Yang, R. and Berger, J.O. (1997). A catalogue of noninformative priors. *Working Paper*, **97-42**, Institute of Statistics and Decision Sciences, Duke University. Available from: .

6. Implementation of Bayesian inference

Objective

To introduce the main numerical methods that can be used to evaluate the integrals necessary in many Bayesian problems. In particular, we concentrate on MCMC and Gibbs sampling approaches.

Recommended reading

- Wiper, M.P. (2007). Introduction to Markov chain Monte Carlo simulation. In *Encyclopedia of Statistics in Quality and Reliability*, Wiley, pp 1014–1020.

Introduction

We have seen that numerical procedures are often needed in Bayesian inference for the computation of the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

and for the computation of posterior moments, predictive distributions etc. The different techniques which might be applied are as follows:

- Numerical integration,
- Gaussian approximations (considered in chapter 8),
- Monte Carlo approaches:
  - ◇ direct methods,
  - ◇ via Markov chains.

Numerical integration

Many numerical integration techniques have been developed. See for example Ausín (2007) or the Wikipedia

http://en.wikipedia.org/wiki/Numerical_integration

for fuller reviews.

One of the simplest approaches is *Simpson's rule*. Supposing that we wish to evaluate the (one dimensional) integral

$$I = \int_a^b g(x)\, dx,$$

in its most simple form, Simpson's rule suggests approximating the integral using

$$I \approx \frac{b-a}{6}\left[g(a) + 4g\left(\frac{a+b}{2}\right) + g(b)\right].$$

This approximation can be improved by subdividing the interval $[a, b]$ into an even number, say N, subintervals

$$[a, a+h) \cup \cdots \cup [a+(N-1)h, a+Nh = b].$$

Using Simpson's rule in each subinterval $[a+jh, a+(j+2)h)$ leads to the final estimate

$$\begin{aligned} I \quad \approx \quad & \frac{h}{3}\left[g(a) + 4g(a+h) + 2g(a+2h) + \cdots\right. \\ & +2g(a+(N-2)h) + 4g(a+(N-1)h) + \\ & \left.+g(a+Nh)\right]. \end{aligned}$$

Example 47

Suppose that we wish to estimate the constant of a beta density, $X \sim \mathcal{B}(7, 10)$, with density function

$$\pi(x) \propto x^6(1-x)^9 \quad \text{for } 0 < x < 1.$$

We shall try to estimate the beta function, $B(7,10) = \int_0^1 x^6(1-x)^9 dx$, using Simpson's rule. Setting $h = 0.1$, we find the following table:

θ	$\theta^6(1-\theta)^9$	$\int_0^\theta \phi^6(1-\phi)^9\, d\phi$
.0	$0.00000E-00$	$0.00000E-00$
.1	$3.87420E-07$	
.2	$8.58994E-06$	$3.37987E-07$
.3	$2.94178E-05$	
.4	$4.12783E-05$	$5.92263E-06$
.5	$3.05176E-05$	
.6	$1.22306E-05$	$1.17753E-05$
.7	$2.31568E-06$	
.8	$1.34218E-07$	$1.24962E-05$
.9	$5.31441E-10$	
1.0	$0.00000E-00$	$1.25007E-05$

The true value of the integral is $B(7,10) = 1.24875E-05$. Using Simpson's rule with $h = 0.05$ gives the result $1.24876E-05$.

An improvement on Simpson's basic rule is the adaptive Simpson's rule, which does not fix the number of subintervals a priori, but instead, continues to subdivide the intervals until the estimated error reaches a given tolerance.

Alternative approaches and problems

Other rules have been designed which take into account the form of the integrand. For example, Gaussian quadrature approaches use an approximation

$$I = \int_a^b g(x)\, dx \approx \sum_{i=1}^{N} w_i g(x_i)$$

where the points x_i are determined as the roots of a class of *orthogonal polynomials*.

The main problem with numerical integration approaches is the curse of dimensionality. As the dimension of the integral increases, the number of function evaluations necessary to achieve a given tolerance increases very rapidly. Thus, in general, such methods are not employed for higher than two or three dimensional integrals.

Monte Carlo approaches

We have seen the basic Monte Carlo method earlier in chapter 3. Suppose that we have $X \sim \pi$ and that we wish to estimate the mean of some functional $E[g(X)]$. Then given a sample, $\mathbf{x}$, of size n from π, we can estimate

$$\bar{g}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} g(x_i) \approx E[g(X)].$$

When $E[g^2(X)]$ exists, then we can estimate the sample variance using

$$V[g(\mathbf{X})] = \frac{1}{N}\int (g(x) - E[g(X)])^2 \approx \frac{1}{N^2}\sum_{i=1}^{N}(g(x_i) - \bar{g}(\mathbf{x}))^2.$$

In many cases however, there is no straightforward way of generating a sample directly from π. In such cases, two main alternatives have been considered: importance sampling and rejection sampling.

Importance sampling

Suppose that sampling from π is complicated. Suppose instead that we can easily sample from another density, say f. Now we can write the expected value of $g(X)$ (under π) as

$$\begin{aligned} E_\pi[g(X)] &= \int g(x)\pi(x)\,dx \\ &= \int \frac{g(x)\pi(x)}{f(x)} f(x)\,dx \\ &= E_f[w(X)g(X)] \end{aligned}$$

where $w(X) = \frac{\pi(X)}{f(X)}$. Thus, we can approximate the expectation by generating a sample of size N from f and using

$$E_\pi[g(X)] \approx \frac{1}{N}\sum_{i=1}^{N} w(x_i)g(x_i) \quad \text{where } w(x_i) = \frac{\pi(x_i)}{f(x_i)} \text{ for } i = 1, \ldots, N.$$

Furthermore, if the density π is known only up to an integration constant, C, then we can extend the approximation to give

$$E_\pi[g(X)] \approx \frac{\sum_{i=1}^N w(x_i)g(x_i)}{\sum_{i=1}^N w(x_i)}$$

where the denominator (divided by N) is an approximation of C.

In general, the choice of *importance function*, f, will strongly influence the efficiency of this algorithm. One should first note that the variance of the importance sampling estimator of $E[g(X)]$ is finite only when the expectation

$$E_f\left[w(X)^2 g(X)^2\right] = E_\pi\left[g(X)^2 w(X)\right] = \int g(x)^2 \frac{\pi(x)^2}{f(x)}\, dx < \infty.$$

This implies that we cannot choose importance functions with lighter tails than π. In the Bayesian context, where we often wish to estimate various posterior expectations, then an efficient importance function will be similar to the true posterior, but with heavier tails.

Example 48

Consider Example 47 where we have $X \sim \mathcal{B}(7, 10)$ so that $\pi(x) \propto x^6(1-x)^9$, and suppose that we wish to estimate the beta function $B(7, 10)$ and the posterior mean

$$E[X] = \frac{1}{B(7,10)} \int_0^1 x^7(1-x)^9 \, dx.$$

One possibility is to use importance sampling with a uniform importance function. In this case, we have importance weights

$$w(x) = \frac{\pi(x)}{1} = x^6(1-x)^9$$

and given a uniform sample of size N, we can estimate $\sum_{i=1}^N w(x_i) \approx B(7, 10)$ and $\frac{\sum_{i=1}^N x_i w(x_i)}{\sum_{i=1}^N w(x_i)} \approx E[X]$.

This is easily programmed in Matlab

```
alpha=7; beta=10; n=1000;
x=rand(1,n);
w=(alpha-1)*log(x)+(beta-1)*log(1-x);
w=exp(w);
betafunctn=sum(w);
w=w/sum(w);
meanx=sum(w.*x);
```

Given an importance sample of size $N = 1000$, the beta function was estimated to be $1.2779E-005$ (true value $1.24875E-005$) and the posterior mean was estimated at 0.4044 (true mean 0.4118). In this example, sample sizes of over 100000 are needed to achieve more than 3 figure accuracy.

Problems

One problem with this approach is that if the importance function is not similar to π (or $|g| \times \pi$) so that the centre of π (or $|g| \times \pi$) is in the tail of the importance function, then this can lead to many of the importance weights being very small and thus, the integral estimates may be largely determined by a very small number of data.

A second problem is that the importance sampling method does not provide a sample from π. This can be remedied by using sampling importance resampling (Rubin 1987).

The SIR algorithm

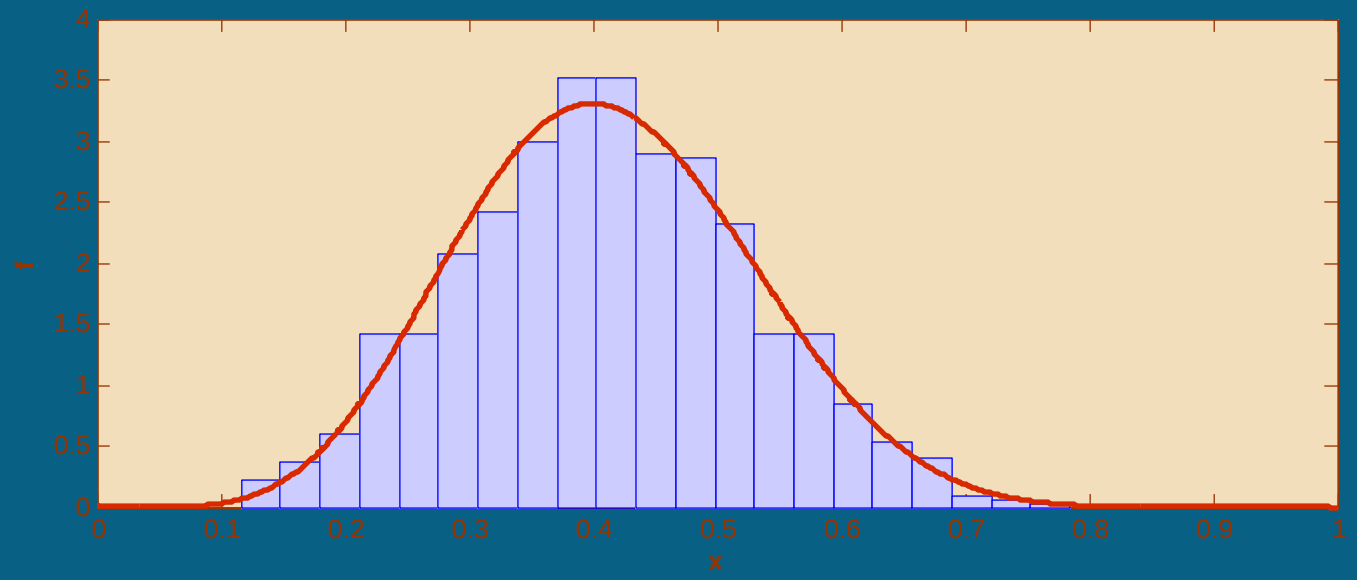
One way of obtaining an approximate sample from π is by subsampling.

If the weights, $w(x_i)$ are normalized so that we define

$$w_i = \frac{w(x_i)}{\sum_{i=1}^{N} w(x_i)}$$

then we can generate an approximate sample, $\tilde{\mathbf{x}}$, of size $M < N$ from π by setting $\tilde{x}_j = x_i$ with probability w_i for $i = 1, \ldots, N$ and $j = 1, \ldots, M$.

The following diagram shows the data simulated using a resample of size 1000 from an importance sample of size 10000.



The sampled data well approximate the beta density.

The rejection algorithm

As with standard Monte Carlo, we assume we wish to generate a sample from $\pi(x)$, which is known only up to a constant. Then the rejection approach chooses to generate data from a proposal distribution, $h(x)$, such that

$$\pi(x) < Mh(x)$$

for some given $M > 0$. The algorithm proceeds as follows.

For $i = 1, \ldots, N$:

1. Generate $\tilde{x}_i \sim h$,
2. Generate $u_i \sim \mathcal{U}(0, 1)$,
3. If $Mu_i h(x_i) < \pi(\tilde{x}_i)$ set $x_i = \tilde{x}_i$.
4. Otherwise, repeat from step 1.

Proof that this algorithm generates a sample from π

Consider $P(X < c)$, where X is generated from this algorithm. We have:

$$
\begin{aligned}
P(X \le c) &= P\left(\tilde{X} \le c | U < \pi(\tilde{X})/(Mh(\tilde{X}))\right) \quad \text{where } \tilde{X} \sim h \text{ and } U \sim \mathcal{U}(0,1) \\
&= \frac{P\left(\tilde{X} \le c, U < \pi(\tilde{X})/(Mh(\tilde{X}))\right)}{P\left(U < \pi(\tilde{X})/(Mh(\tilde{X}))\right)} \\
&= \frac{\int_{-\infty}^{c} \int_{0}^{\pi(\tilde{x})/(Mh(\tilde{x}))} h(\tilde{x})\, du\, d\tilde{x}}{\int_{-\infty}^{\infty} \int_{0}^{\pi(\tilde{x})/(Mh(\tilde{x}))} h(\tilde{x})\, du\, d\tilde{x}} \\
&= \frac{\int_{-\infty}^{c} \frac{\pi(\tilde{x})}{Mh(\tilde{x})} h(\tilde{x})\, d\tilde{x}}{\int_{-\infty}^{\infty} \frac{\pi(\tilde{x})}{Mh(\tilde{x})} h(\tilde{x})\, d\tilde{x}} \\
&= \frac{\int_{-\infty}^{c} \pi(\tilde{x})\, dx}{\int_{-\infty}^{\infty} \pi(\tilde{x})\, dx} = P(X < c) \text{ where } X \sim \pi.
\end{aligned}
$$

∎

This algorithm clearly reduces to standard Monte Carlo sampling when $h = \pi$. Otherwise, as with importance sampling, it is necessary that the tails of the proposal distribution are thicker than those of π.

The main problem with this approach is finding a good proposal distribution so that only a small number of candidates are rejected. Note that the probability of accepting a draw (assuming that π is properly scaled to integrate to 1) is

$$\begin{aligned} P\left(U < \frac{\pi(\tilde{X})}{Mh(\tilde{X})}\right) &= \int_{-\infty}^{\infty} \int_{0}^{\pi(\tilde{x})/(Mh(\tilde{x}))} h(\tilde{x})\, du\, d\tilde{x} \\ &= \int_{-\infty}^{\infty} \frac{\pi(\tilde{x})}{Mh(\tilde{x})} h(\tilde{x})\, d\tilde{x} \\ &= \frac{1}{M} \end{aligned}$$

so that we would like M to be as close to 1 as possible.

Example 49

Suppose that we wish to simulate from a truncated normal distribution $X \sim \mathcal{TN}(0,1)$ where $X > \alpha > 0$. One way to do this would be to sample directly from the $\mathcal{N}(0,1)$ density and simply reject those values that fall below α. However, this method could be very inefficient if α is large. In this case, an alternative is to use a shifted, exponential distribution

$$h(x) = \lambda e^{-\lambda(x-\alpha)} \quad \text{for } x > \alpha$$

as an envelope function. (More sophisticated algorithms are proposed by Geweke (1991) and Robert (1995).)

Writing the normal density as $\pi(x) = ce^{-\frac{x^2}{2}}$, where $\frac{1}{c} = \sqrt{2\pi}\,(1 - \Phi(\alpha))$, then,

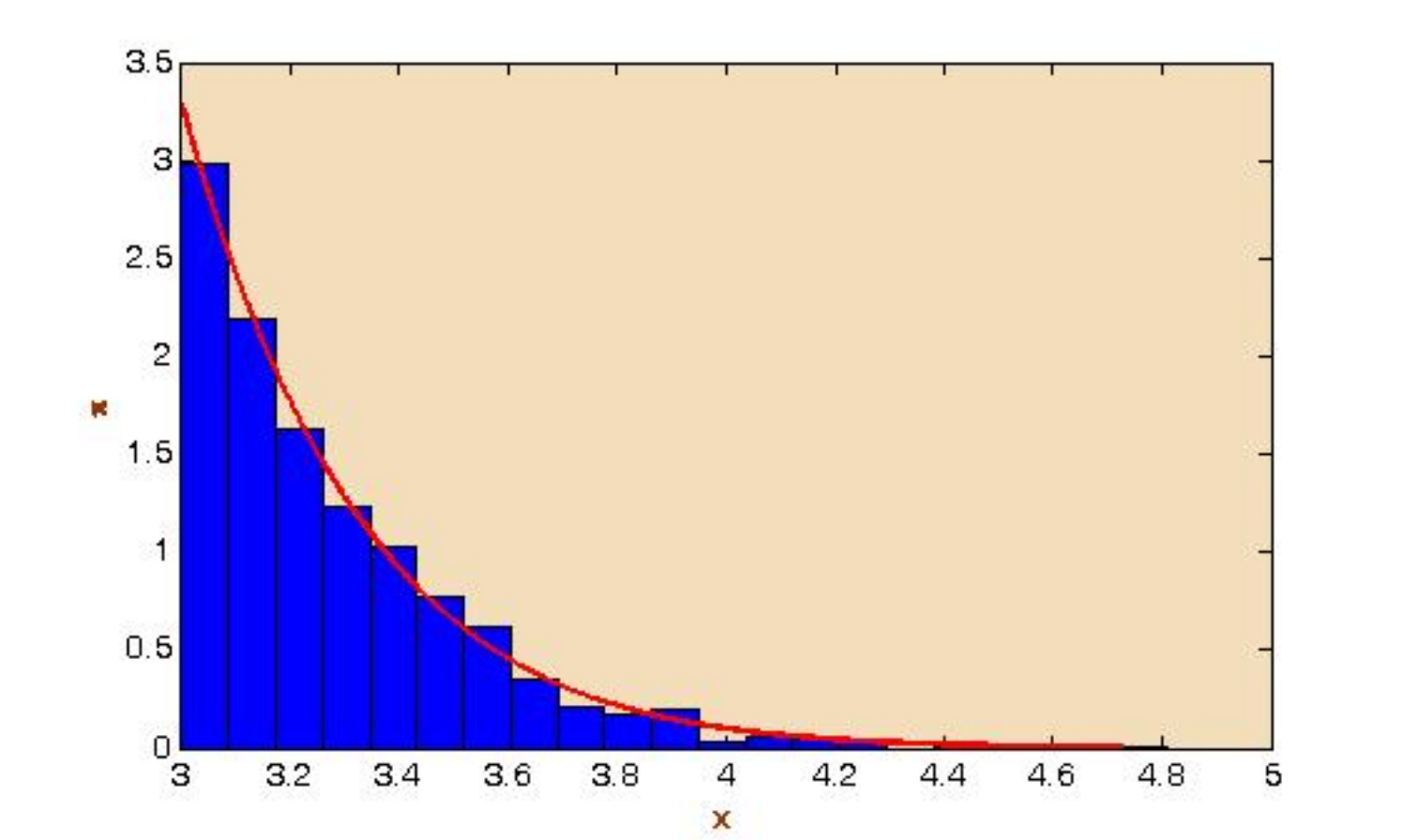
$$\begin{aligned}
\frac{\pi(x)}{h(x)} &= \frac{c}{\lambda}\exp\left(-\frac{x^2}{2} + \lambda(x-\alpha)\right) \quad \text{for } x > \alpha \\
&\leq \frac{c}{\lambda}\exp\left(\max_{x>\alpha}\left[-\frac{x^2}{2} + \lambda(x-\alpha)\right]\right) \\
&= M_1(\lambda)I(\lambda > \alpha) + M_2(\lambda)I(\lambda \leq \alpha) \quad \text{where } I \text{ is the indicator and} \\
M_1(\lambda) &= \frac{c}{\lambda}\exp\left(\frac{\lambda^2}{2} - \alpha\lambda\right) \\
M_2(\lambda) &= \frac{c}{\lambda}\exp\left(-\frac{\alpha^2}{2}\right)
\end{aligned}$$

To get the most efficient routine we should choose λ to minimize this function. However, it is simpler to minimize only M_2. This gives $\lambda = \alpha$.

In this case, the probability that a generated candidate, $\tilde{x}$ is accepted is

$$\frac{\pi(\tilde{x})}{h(\tilde{x})M_2(\alpha)} = \exp\left(-\frac{1}{2}(\tilde{x}^2 + \alpha^2) + \alpha\tilde{x}\right).$$

The following is the fitted distribution when $\alpha = 3$. Only 86 out of 1000 proposed values were rejected. We can see that the fit is very good. The probability of accepting a value generated from an untruncated normal distribution is only 0.0013.


3.5
3
2.5
2
1.5
1
0.5
0
3
3.2
3.4
3.6
3.8
4
4.2
4.4
4.6
4.8
5
x

Envelope methods

These are refinements of the basic algorithm based on bounding the target density from above and below. Suppose that we can find a proposal density h and a (non-negative) function g such that

$$g(x) \leq \pi(x) \leq Mh(x) \quad \text{for all } x.$$

Then, the following algorithm generates a variable, X, with distribution π.

1. Generate $\tilde{X} \sim h$ and $U \sim \mathcal{U}(0,1)$.

2. If $U \leq \frac{g(\tilde{X})}{Mh(\tilde{X})}$ let $X = \tilde{X}$.

3. Otherwise, let $\tilde{X} = X$ if $U \leq \frac{\pi(\tilde{X})}{Mh(\tilde{X})}$

4. Otherwise, repeat from step 1.

The advantage of this algorithm is that the number of necessary evaluations of π are reduced, and instead, we often only need to evaluate the (simpler) densities, g and h. The probability that π does not have to be evaluated is $\frac{1}{M}\int g(x)\,dx$ which reflects the potential gain in using this approach.

One particular case that allows for the simple construction of bounding functions is when the density, π, is $log\ concave$.

Definition 11
A density $f(x)$ is said to be log concave if $\frac{\partial^2}{\partial x^2} f(x) < 0 \,\forall\, x$.

Most exponential family densities are log-concave. For example, if $X|\theta \sim \mathcal{N}(\theta, 1)$, then $\frac{\partial^2}{\partial x^2} f(x|\theta) = -1$.

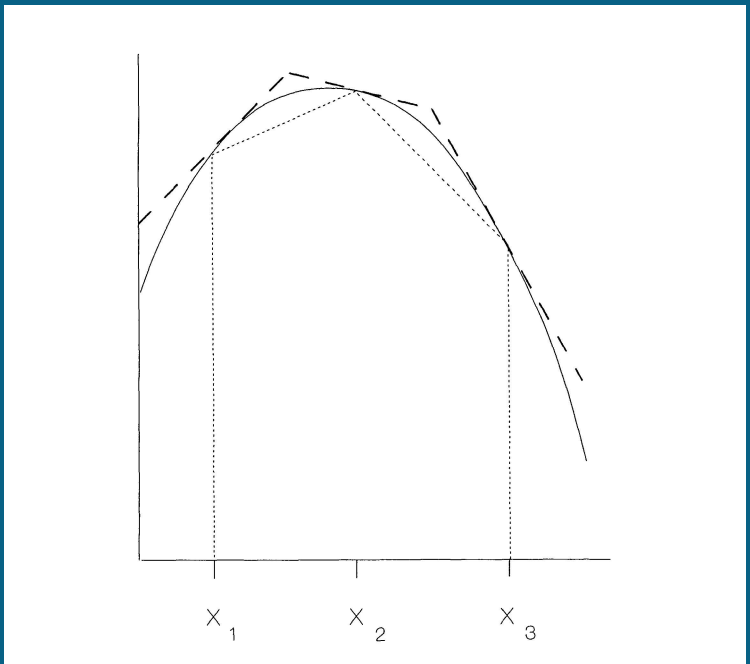
Adaptive rejection sampling

This algorithm developed by Gilks (1992) and Gilks and Wild (1992) gives a general method of constructing the bounding functions g and h when the target density, π, is log concave.

Suppose that S_n is a set of points x_i for $i = 0, 1, \ldots, n+1$ in the support of π such that $\log \pi(x_i)$ is known up to the same constant. As log π is concave, then the line $L_{i,i+1}$ going through $(x_i, \log \pi(x_i))$ and $(x_{i+1}, \log \pi(x_{i+1}))$ lies below the graph of $\log \pi$ in $(x_i, x_{i+1}]$ and lies above the graph outside this interval. Thus, for the interval $(x_i, x_{i+1}]$, we can define $\bar{\phi}_n(x) = \min L_{i-1,i}(x), L_{i+1,i+2}(x)$ and $\underline{\phi}_n(x) = L_{i,i+1}(x)$ which bound $\log \pi$. Defining $H_n(x) = \exp(\bar{\phi}_n(x))$ and $g_n(x) = \exp(\underline{\phi}_n(x))$, we have

$$g_n(x) \leq \pi(x) \leq H_n(x) = M_n h_n(x) \text{say}$$

where h_n is a density function.



An advantage of this approach is that if a generated value is rejected, it can then be added to the set S_n which improves the bounds on π at the next step. This leads to the following generic algorithm.

The ARS algorithm

1. Initialize n and S_n.

2. Generate $\tilde{X} \sim h_n$ and $U \sim \mathcal{U}(0,1)$.

3. If $U < g_n(\tilde{X})/h_n(\tilde{X})$ then set $X = \tilde{X}$.

4. Otherwise, if $U < \pi(\tilde{X})/(M_n h_n(\tilde{X})$, set $X = \tilde{X}$.

5. Otherwise, set $n = n+1$, $S_{n+1} = S_n \cup \tilde{X}$ and repeat from 2.

The big advantage of this algorithm is its universality. As long as π is known to be log concave, it can always be used.

MCMC methods

As simple Monte Carlo algorithms are not always straightforward to implement, another alternative is to use algorithms which generate approximate Monte Carlo samples. The most popular approach is Markov chain Monte Carlo or MCMC which samples from a Markov chain whose limit distribution is the distribution from which we wish to sample.

Markov chains

Definition 12

A *Markov chain*, $\{X_t\}$, is defined to be a sequence of variables, $X_0, X_1, X_2, \ldots$, such that the distribution of X_t given the previous values $X_0, \ldots, X_{t-1}$ only depends on X_{t-1}, so that

$$P(X_t \in A | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = P(X_t \in A | X_{t-1} = x_{t-1})$$

for all $A, x_1, \ldots, x_{t-1}$.

Most Markov chains that we deal with are *time-homogeneous*, that is

$$P(X_{t+k} \in A | X_t = x) = P(X_k \in A | X_0 = x) \quad \text{for any } k.$$

A simple example of a time-homogeneous Markov chain is a random walk

$$X_t = X_{t-1} + \epsilon_t \quad \text{where } \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$$

It is clear that a time-homogeneous Markov chain is completely defined by the initial state, X_0, and by the *transition kernel*,

$$P(x, y) = P(X_{t+1} = y | X_t = x).$$

For most problems of interest, the Markov chain will take values in a continuous, multivariate state space. However, we shall assume initially that the state space is finite and countable, so that we can assume that $X_t \in \{1, 2, \ldots, k\}$ for some k.

In this case, we can define the *t-step transition probabilities*

$$p_{ij}(t) = P(X_t = j | X_0 = i)$$

and then, we can consider the conditions under which these probabilities converge, i.e. that

$$p_{ij}(t) \to \pi(j) \quad \text{as } t \to \infty.$$

Definition 13
A Markov chain is said to be $irreducible$ if for every i, j, there exists some t such that $p_{ij}(t) > 0$.

Irreducibility implies that it is possible to visit every state in the chain starting from any initial state.

Definition 14
A state, i, of a Markov chain is said to be $recurrent$ if return to state i is certain, that is if we define τ_i to be the number of steps needed to return to state i, then $P(\tau_i < \infty) = 1$. A recurrent state is further said to be $positive\ recurrent$ if the expected return time is finite, so that $E[\tau_i] < \infty$.

Definition 15
The $period$ of a state, i, is defined to be $d(i) = \gcd\{t : p_{ii}(t) > 0\}$. A state with period 1 is said to be $aperiodic$.

It can be shown that if any state of an irreducible chain is positive recurrent, then all states are and also that all states in such a chain have the same period.

The equilibrium distribution of a Markov chain

Theorem 28

For an irreducible, positive definite, aperiodic Markov chain with t step transition density $p_{ij}(t)$, then a unique equilibrium distribution π exists so that for all i, j,

$$\pi(j) = \lim_{t \to \infty} p_{ij}(t).$$

Proof ∎

It can be shown that a sufficient condition for the existence of a unique stationary distribution is *reversibility*. A Markov chain with transition probabilities $p_{ij} = P(X_{t+1} = j | X_t = i)$ is said to be reversible if there exists a probability density π that satisfies *detailed balance*, so that for any i, j, then

$$p_{ij}\pi(i) = p_{ji}\pi(j).$$

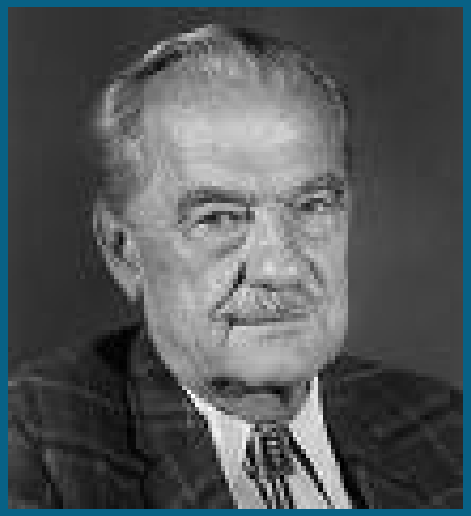
Markov chains with continuous state space

It is possible to extend the previous arguments to Markov chains with a continuous state space, although the conditions for the equilibrium distribution are slightly more technical, see e.g. Robert and Casella (2004). In this case, given a transition kernel, $P(x, y)$, then a stationary distribution π must satisfy

$$\pi(y) = \int P(x, y)\pi(x)\, dx.$$

From a Bayesian viewpoint, the objective of the MCMC approach is thus to construct a Markov chain with a given stationary distribution π which is the Bayesian posterior distribution.

The Metropolis Hastings algorithm



Metropolis

This is a general algorithm for constructing a Markov chain and was introduced by Metropolis et al (1953) and extended by Hastings (1970). The general algorithm for generating a chain with equilibrium distribution π is as follows:

The algorithm

1. Given the current value, $X_t = x$, generate a candidate value, y, from a proposal density $q(y|x)$.

2. Calculate the acceptance probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}.$$

3. With probability $\alpha(x, y)$ define $X_{t+1} = y$ and otherwise reject the proposed value and set $X_{t+1} = x$.

4. Repeat until convergence is judged and a sample of the desired size is obtained.

Why does this algorithm work?

The transition kernel of a move from x to y is

$$P(x,y) = \alpha(x,y)q(y|x) + \left(1 - \int \alpha(x,y)q(y|x)\,dy\right)\delta_x$$

where δ_x is the Dirac delta function at x.

Now, it is easy to show that

$$\pi(x)q(y|x)\alpha(x,y) = \pi(y)q(x|y)\alpha(y,x)$$

and that

$$\left(1 - \int \alpha(x,y)q(y|x)\,dy\right)\delta_x = \left(1 - \int \alpha(y,x)q(x|y)\,dx\right)\delta_y.$$

This implies that we have detailed balance

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

so that π is a stationary distribution of the chain. ∎

It is important to notice that the Metropolis Hastings acceptance probability only depends on pi through the ratio $\pi(y)/\pi(x)$. This is particularly useful in the Bayesian context, when the form of the posterior distribution is usually known up to the constant of integration.

Also note that when the proposal density $q(y|x) = \pi(y)$, then the Metropolis Hastings acceptance probability is exactly 1 and the algorithm is the same as standard Monte Carlo sampling.

Choosing a proposal density

One might expect that the Metropolis Hastings algorithm would be more efficient if $\alpha(x, y)$ was high. Unfortunately, this is not usually the case. In Roberts et al (1997), it is recommended that for high dimensional models, the acceptance rate for random-walk algorithms (see later) should be around 25% whereas in models of dimension 1 or 2, this should be around 50%.

However, general results are not available and the efficiency of a Metropolis Hastings algorithm is usually heavily dependent on the proposal density $q(y|x)$.

The independence and Metropolis samplers

The *independence sampler* defines a proposal density $q(y|x) = q(y)$ independent of x. This will often work well if the density q is similar to π, although with somewhat heavier tails, similarly to the Monte Carlo rejection sampler.

Another alternative is the *Metropolis* (1953), sampler which has the property that $q(x|y) = q(y|x)$. One small advantage of this approach is that the acceptance probability simplifies down to the

$$\alpha(x, y) = \frac{\pi(y)}{\pi(x)}.$$

A special case is the *random walk Metropolis* algorithm which assumes that $q(y|x) = q(|y - x|)$. For example, in univariate problems, one might consider a normal proposal density $q(y|x) = \mathcal{N}\left(x, \sigma^2\right)$ where the value of σ can be adjusted to achieve an acceptable acceptance rate.

Example

Example 50

Suppose that we observe a sample of size n from a Cauchy distribution, $X|\theta \sim \mathcal{C}(\theta, 1)$, that is

$$f(x|\theta) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} \quad \text{for } -\infty < x < \infty$$

Given a uniform prior for θ, then the posterior distribution is

$$p(\theta|\mathbf{x}) \propto \prod_{i=1}^{n} \frac{1}{1 + (x_i - \theta)^2}.$$

One way of sampling this distribution is to use a random walk Metropolis algorithm. We could use a Cauchy proposal density, $\tilde{\theta}|\theta \sim \mathcal{C}(\theta, \sigma)$, so that

$$q(\tilde{\theta}|\theta) = \frac{1}{\pi\sigma\left(1 + \left(\frac{\tilde{\theta}-\theta}{\sigma}\right)^2\right)}.$$

The scale parameter, σ, can be adjusted to achieve the desired acceptance rate.

In this case, the probability of accepting a proposed value, $\tilde{\theta}$ given the current value, θ, is

$$\alpha(\theta, \tilde{\theta}) = \min\left\{1, \prod_{i=1}^{n} \frac{1 + (x_i - \theta)^2}{1 + (x_i - \tilde{\theta})^2}\right\}.$$

As an alternative, an independence sampler could be proposed. In this case, we might assume a Cauchy proposal distribution, $\tilde{\theta} \sim \mathcal{C}(m, \tau)$, where the location parameter, m, is the sample median. In this case, the acceptance probability is

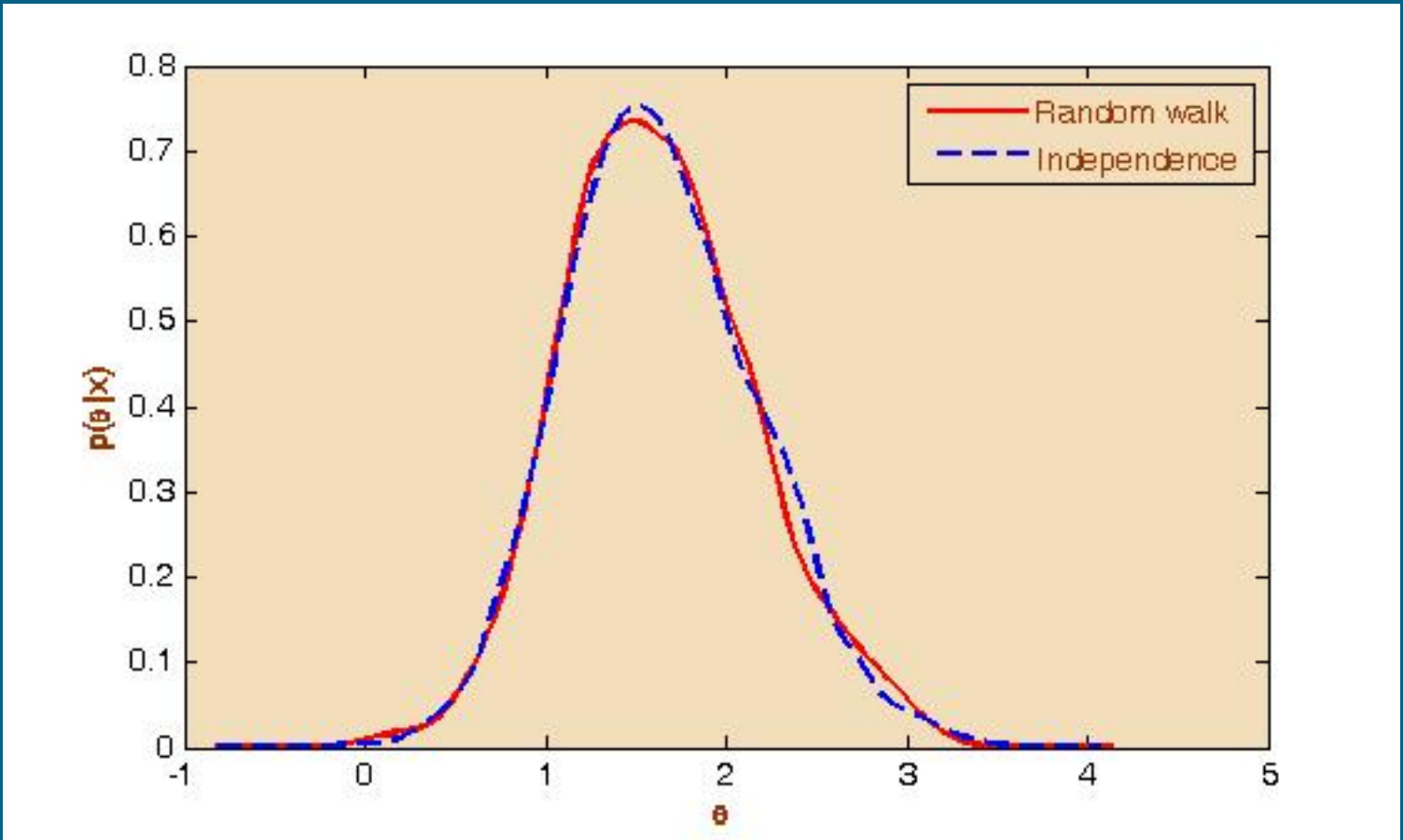
$$\min\left\{1, \prod_{i=1}^{n} \frac{1 + (x_i - \theta)^2}{1 + (x_i - \tilde{\theta})^2} \frac{1 + \left(\frac{\tilde{\theta} - m}{\tau}\right)^2}{1 + \left(\frac{\theta - m}{\tau}\right)^2}\right\}.$$

A sample of 10 data were generated from a Cauchy distribution, $X \sim \mathcal{C}(1, 1)$, with the following results:

$$\mathbf{x} = \begin{matrix} -5.1091 & -0.7651 & 0.9261 & 1.0232 & 1.1669 \\ 1.2702 & 2.4846 & 2.5375 & 3.3476 & 3.6066 \end{matrix}$$

Both the random walk sampler (with $\sigma = 0.3$) and the independence sampler (with $\tau = 0.5$) were run for 10000 iterations, starting from the sample median. For the random walk sampler, 66.7% of the proposed values were accepted and for the independence sampler, around 52% of the proposals were accepted. The samplers took a few seconds to run in each case.

Kernel density estimates of the posterior density of θ given a uniform prior are given in the following diagram.



The estimated density function is approximately the same in each case.

Block Metropolis Hastings

When the dimension of X is large, then it can often be difficult to find a reasonable proposal density. In this case, it is sensible to divide X into blocks, say $X = (X_1, \ldots, X_k)$ and construct a chain with these smaller blocks.

Suppose initially that $X = (X_1, X_2)$ and define two proposal densities $q_1(y_1|x_1, x_2)$, $q_2(y_2|x_1, x_2)$ to generate candidate values for each component.

Then, define the acceptance probabilities

$$\begin{aligned} \alpha_1(x_1, y_1|x_2) &= \min\left\{1, \frac{\pi(y_1|x_2)q_1(x_1|y_1, x_2)}{\pi(x_1|x_2)q_1(y_1|x_1, x_2)}\right\} \\ \alpha_2(x_2, y_2|x_1) &= \min\left\{1, \frac{\pi(y_2|x_1)q_2(x_2|x_1, y_2)}{\pi(x_2|x_1)q_2(y_2|x_1, x_2)}\right\} \end{aligned}$$

where the densities $\pi(x_1|x_2)$ and $\pi(x_2|x_1)$ are the conditional densities and $\pi(x_1|x_2) \propto \pi(x_1, x_2)$.

The algorithm now proceeds by successively sampling from each block in turn.

The slice sampler

The slice sampler (Neal 2003) is an attractive approach when the state space is relatively low dimensional. The general algorithm for sampling from π is

1. Given a current value, X_t, simulate $U_{t+1} \sim \mathcal{U}[0, \pi(X_t)]$.

2. Simulate $X_{t+1} \sim \mathcal{U}[A_{t+1}]$, where $A_{t+1} = \{x : \pi(x) \geq U_{t+1}\}$.

It is clearly unimportant whether the constant of integration is known or not.

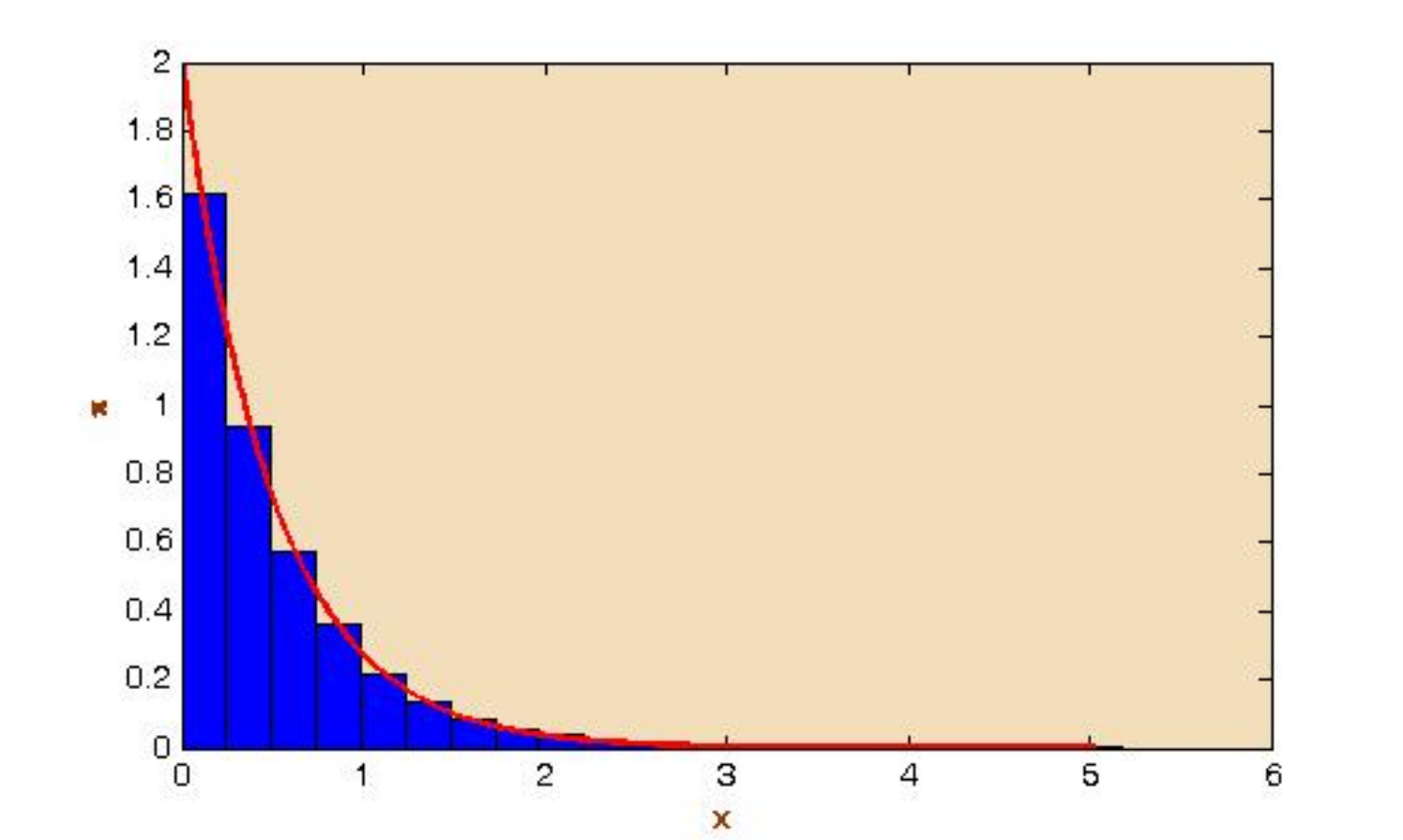
Example

Example 51
Suppose that we wish to sample from an exponential density $X \sim \mathcal{E}(\lambda)$. Then, we know that $\pi(x) \propto e^{-\lambda x}$ and a slice sampler could proceed as follows.

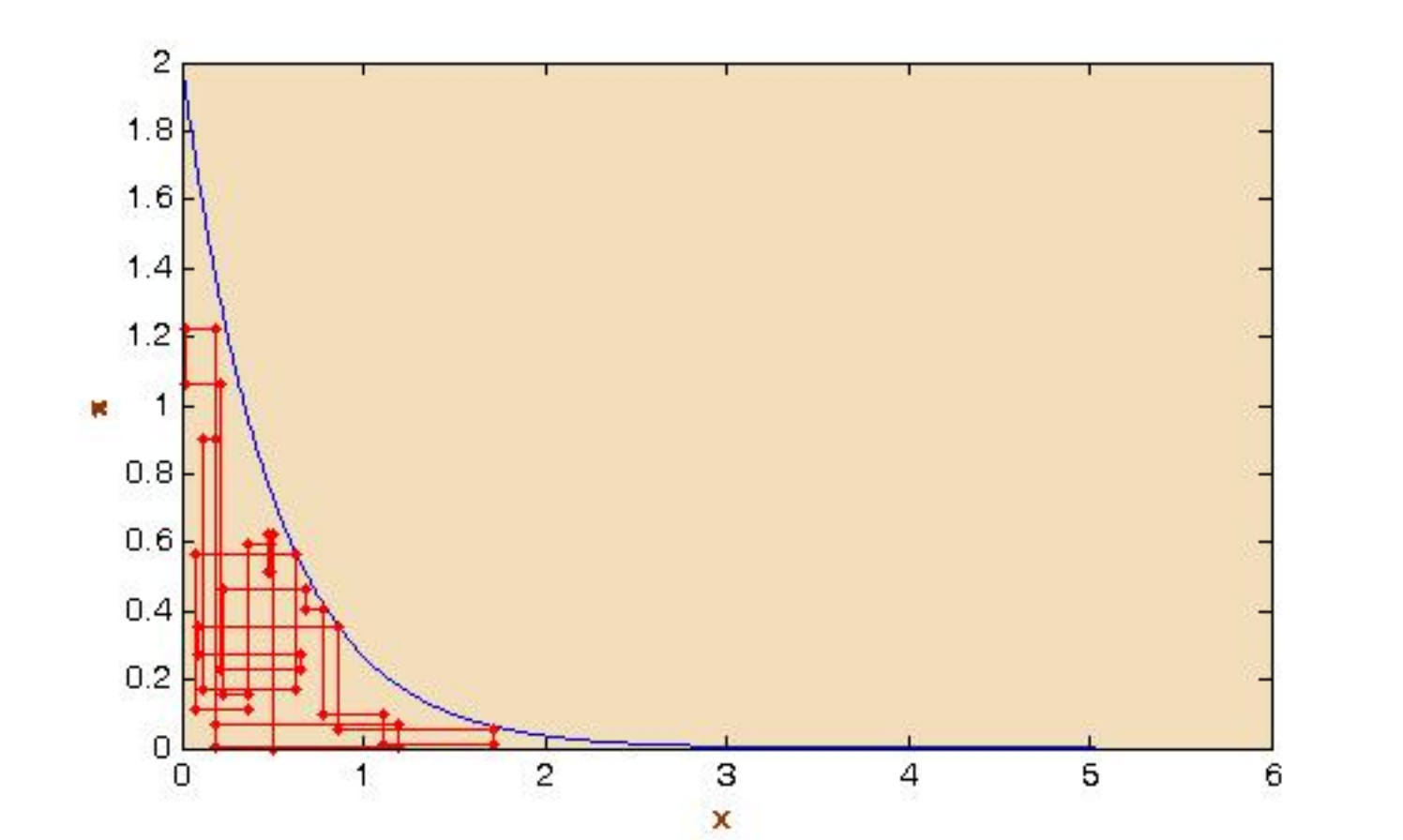
1. Given X_t, generate $U_{t+1} \sim \mathcal{U}\left[0, e^{-\lambda x_t}\right]$.

2. Generate $X_{t+1} \sim \mathcal{U}\left[0, -\frac{1}{\lambda}\log U_{t+1}\right]$.

The following diagram illustrates the results of 10000 iterations when $\lambda = 2$.



The final diagram illustrates how the chain moves.


2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
y
0
1
2
3
4
5
6
x

Gibbs sampling

We have seen Gibbs sampling previously in chapter 3. If we assume that $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$ has joint distribution π and that the conditional distributions $\pi_i(\mathbf{X}_i|\mathbf{X}_{-i})$ are all available, where $\mathbf{X}_{-1} = (\mathbf{x}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_k)$ then the Gibbs sampler generates an (approximate) sample from π by successively sampling from these conditional densities. Thus, assuming that the current values are $\mathbf{x}_t$, then the algorithm is the following

1. Generate $\mathbf{x}_{1,t+1} \sim \pi_1(\cdot|\mathbf{x}_{-1,t})$.

2. Generate $\mathbf{x}_{2,t+1} \sim \pi_2(\cdot|x_{1,t+1}, x_{3,t}, \ldots, x_{k,t})$.

3. $\vdots$

4. Generate $\mathbf{x}_{k,t} \sim \pi_k(\cdot|\mathbf{x}_{-k,t+1})$

We can note that Gibbs sampling is a particular version of block Metropolis Hastings algorithm where the proposal distribution for $\mathbf{X}_i$ is exactly the conditional distribution $\pi_i(\mathbf{X}_i|\mathbf{X}_{-1})$ so that the acceptance probability is always equal to 1.

Gibbs sampling can be applied in a remarkably large number of problems.

Example

Example 52

Suppose that the lifetime (in hours) of a machine, Y, has normal distribution so that $Y|X = x \sim \mathcal{N}(x, 1)$ and that we observe n machines during α hours. If, at the end of this time, n_1 machines have failed, with failure times $y_1, \ldots, y_{n_1}$ and $n_2 = n - n_1$ machines are still working, then the likelihood function is

$$l(x|\mathbf{y}) \propto \exp\left(-\frac{n_1}{2}(x - \bar{y}_1)^2\right) (1 - \Phi(\alpha - x))^{n_2}$$

where $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$. Thus, an explicit form for the posterior of x (supposing a uniform prior) is unavailable.

However, suppose that we knew the true values of the latent variables, $Y_{n_1+1} = y_{n_1+1}, \ldots, Y_n = y_n$. Then it is clear that $X|\mathbf{y} \sim \mathcal{N}\left(\bar{y}, \frac{1}{n}\right)$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Also, for $i = n_1 + 1, \ldots, n$, we have $Y_i | X = x, Y_i > \alpha \sim \mathcal{TN}(x, 1)$ truncated onto the region $Y_i > \alpha$. Therefore, we can set up a simple Gibbs sampling algorithm to estimate the posterior density of x as follows.

1. Set $t = 0$ and fix an initial value x_0.

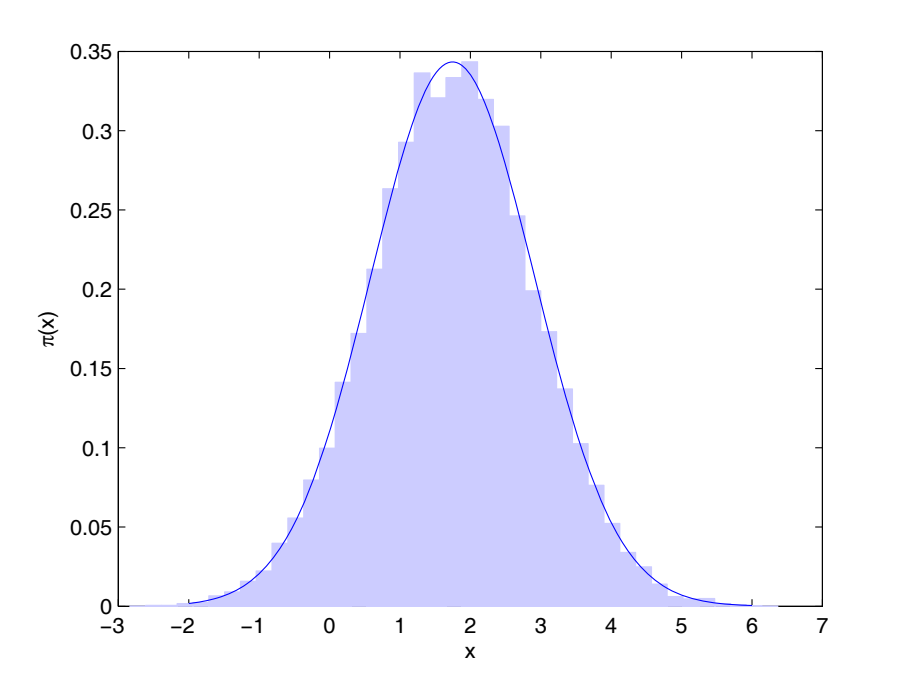
2. For $i = n_1 + 1, \ldots, n$, generate $y_{i,t} \sim \mathcal{TN}(x_t, 1)$.

3. Calculate $\bar{y}_t = \frac{1}{n}\left(\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n} y_{i,t}\right)$.

4. Generate $x_{t+1} \sim \mathcal{N}\left(\bar{y}_t, \frac{1}{n}\right)$.

5. $t = t + 1$. Go to 2.

Wiper (2007) considers a sample of 20 normally distributed lifetimes with mean μ and standard deviation 5, $(Y|X = x \sim \mathcal{N}(x, 25))$, where 14 data less than 5 are observed completely and have mean 0.94 and the remaining 6 data are truncated so that it is only known that they take values greater than 5 and a uniform prior distribution for x is assumed. The following diagram shows a histogram of the values of x generated from 10000 iterations of the Gibbs algorithm along with a fitted density.



Universal implementation of Gibbs sampling algorithms

In many cases, the conditional distributions used in Gibbs samplers are log-concave. In these cases, universal Gibbs samplers can be set up by using the ARS to sample these conditional distributions.

For non log-concave distributions, the adaptive rejection Metropolis sampler (ARMS) was introduced in Gilks et al (1995).

These algorithms form the basis of the `Winbugs` program.

A disadvantage of such universal algorithms is however that they are often inefficient. It is generally preferable to implement specific algorithms tailored to the particular problem.

MCMC convergence assessment

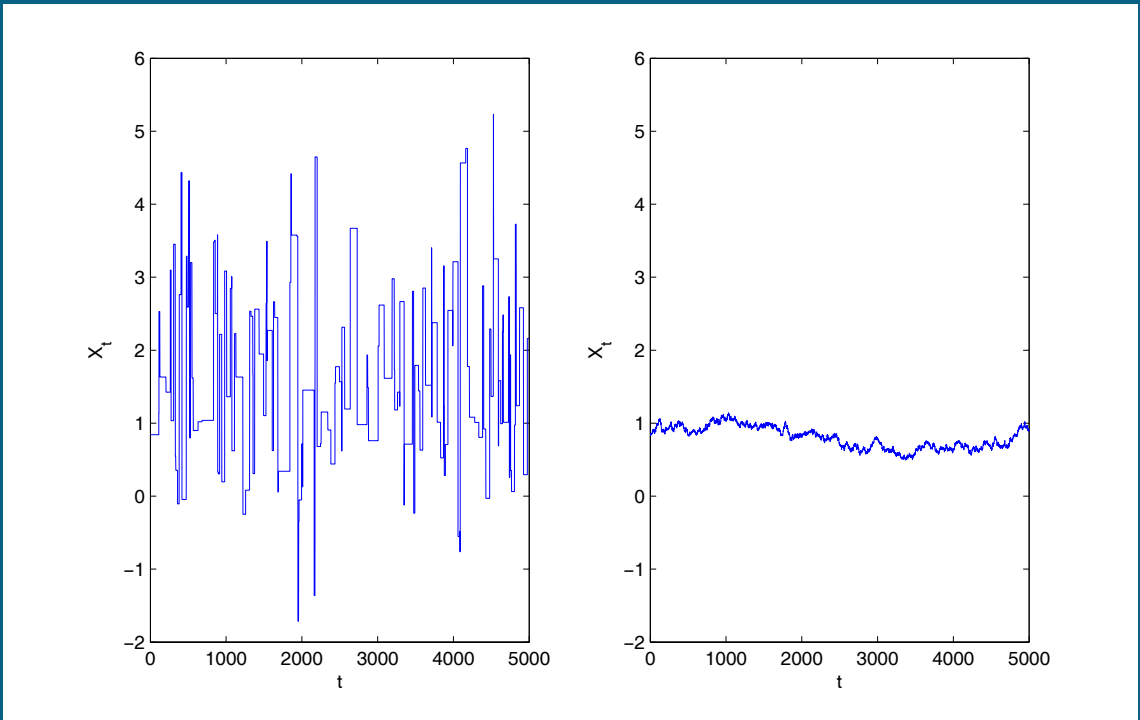
When running an MCMC algorithm, it is important to assess when the sampled values X_t have approximately converged to the stationary distribution π. This will depend on how well the MCMC algorithm is able to explore the state space also on the correlation between the X_t's.

Secondly, we need to assess the convergence of MCMC averages, e.g. $\frac{1}{T}\sum_{t=1}^{T} X_t \rightarrow E[X_t]$ and finally we need to be able to assess how close a given sample is to being independent and identically distributed.

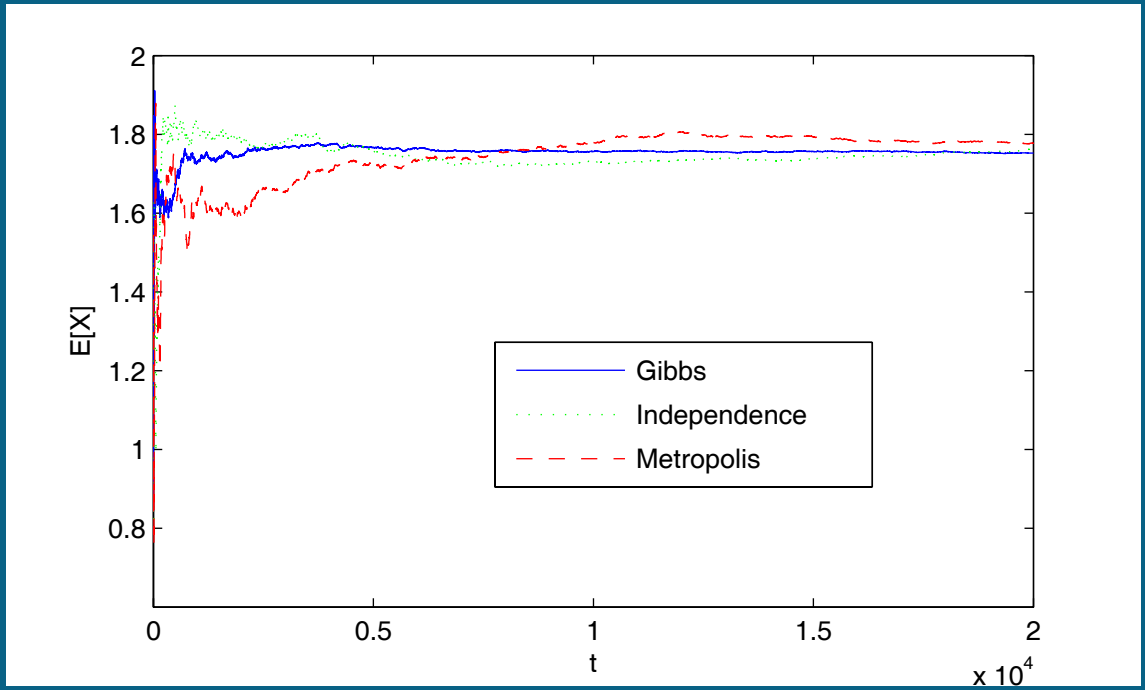
One possibility is to consider running the chain various times with different, disperse starting values. Then, we could assess the convergence of the chain by examining when sample means of the functions of interest generated from each run have converged. Other, formal diagnostics are given in Gelman and Rubin (1992).

The alternative is to use a single run of the chain.

In this case, we can produce graphs of X_t against t to show the mixing of the chain and any deviations from stationarity. The following diagram from Wiper (2007) shows examples of badly mixing chains.

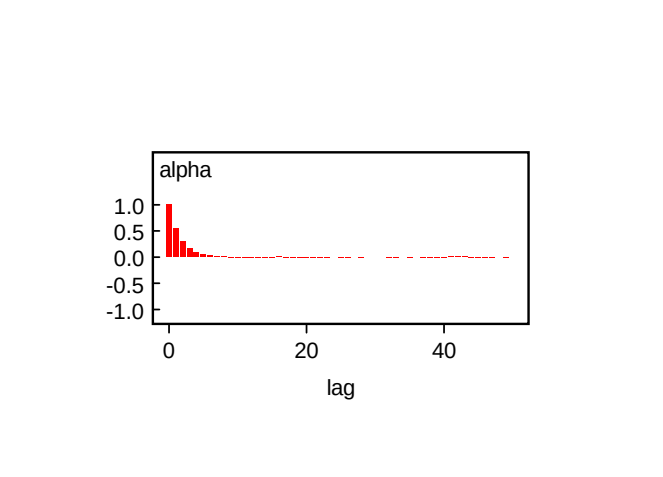


Secondly, we can plot running means of the parameters of interest to see when they have converged. The following diagram shows the estimates of $E[X|\mathbf{y}]$ used from running 3 different algorithms for the problem of Example 52.



It can be seen that the means appear to have converged after about 10000 iterations. Thus, one possibility is to run the sampler for longer, using these initial iterations as a $burn\ in$ period.

thirdly, we can plot the autocorrelation functions of the generated values. In general, as we are generating from a Markov chain, the successive values, X_t, will be positively correlated. Thus, if we wish to estimate, for example, the variance of X, then we must take this correlation into account. The following diagram shows the ACF of the parameter α in the pump failure problem analyzed in Chapter 10.



The autocorrelation has disappeared after about lag 5. One possibility is thus to thin the sample, choosing just every 5th datum which are now, approximately independent.

Other algorithms

Reversible jump

This approach (Green 1995) is basically a Metropolis Hasting sampler, which allows the chain to move over a variably dimensioned model space.

Perfect sampling

This method, developed by Propp and Wilson (1996), uses the idea of coupling from the past in order to generate an exact MCMC sample from π, avoiding the need for converegence diagnostics. See e.g.

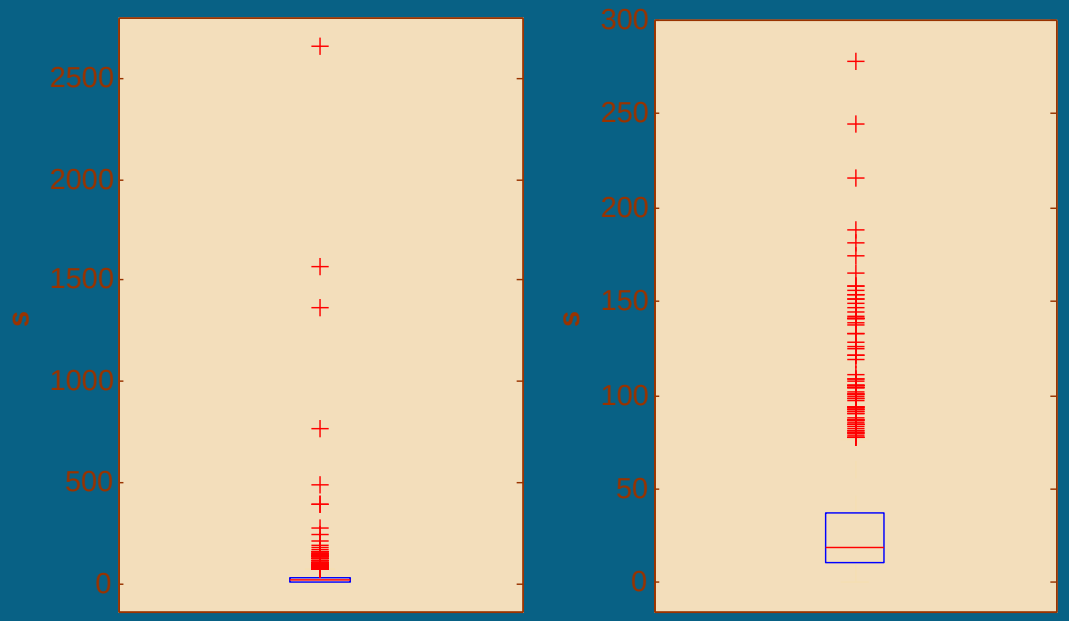
http://dbwilson.com/exact/

Particle filtering

This is an alternative approach to MCMC based on importance sampling and particularly suitable for sequential inference problems. See e.g. Doucet et al (2000) or

http://en.wikipedia.org/wiki/Particle_filter

Application III: Bayesian inference for the $dPlN$ distribution and the $M/G/c/c$ loss system



The boxplots show the times spent (lhs) and times spent below 300 days (rhs) of patients in a geriatric ward of a hospital.

These data have been analyzed previously using Coxian, phase type distributions by e.g. Ausín et al (2003). However, the data are long tailed and therefore, a heavy tailed model should be more appropriate.

Typically, long tailed data are modeled using a Pareto distribution, or a mixture of Pareto distributions (Ramírez et al 2008a) but although such a model can capture the tail behaviour, it does not capture the body of the distribution.

The double Pareto lognormal ($dPlN$) distribution has been recently introduced as a model for heavy tailed data by Reid and Jorgersen (2004).

The skewed Laplace and double Pareto lognormal distributions

It is easiest to define the $dPlN$ distribution by starting from the skewed Laplace distribution.

Definition 16
A random variable, Y is said to have a skewed Laplace distribution with parameters $\mu, \sigma, \alpha, \beta$, that is $Y \sim \mathcal{SL}(\mu, \sigma, \alpha, \beta)$, if

$$f_Y(y) = \frac{\alpha\beta}{\alpha+\beta}\phi\left(\frac{y-\mu}{\sigma}\right)\left[R(\alpha\sigma - (y-\mu)/\sigma) + R(\beta\sigma + (y-\mu)/\sigma)\right]$$

for $y \geq 0$, where $R(y)$ is Mills' ratio, that is

$$R(y) = \frac{1-\Phi(y)}{\phi(y)}.$$

If $Y \sim \mathcal{SL}(\mu, \sigma, \alpha, \beta)$ is a skewed Laplace random variable, then we can write

$$Y = Z + W$$

where $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$, $W = W_1 - W_2$ and $W_1 \sim \mathcal{E}(\alpha)$ and $W_2 \sim \mathcal{E}(\beta)$.

The conditional distributions of $Z|Y = y$ and $W_1|Y = y, Z = z$ are:

$$f_Z(z|y) = p\frac{\frac{1}{\sigma}\phi\left(\frac{z-(\mu-\sigma^2\beta)}{\sigma}\right)}{\Phi^c\left(\frac{y-(\mu-\sigma^2\beta)}{\sigma}\right)}I_{z\geq y} + (1-p)\frac{\frac{1}{\sigma}\phi\left(\frac{z-(\mu+\sigma^2\alpha)}{\sigma}\right)}{\Phi^c\left(\frac{y-(\mu+\sigma^2\alpha)}{\sigma}\right)}I_{z<y} \quad \text{where}$$

$$p = \frac{R(\beta\sigma + (y-\mu)/\sigma)}{R(\alpha\sigma - (y-\mu)/\sigma) + R(\beta\sigma + (y-\mu)/\sigma)} \tag{1}$$

$$f_{W_1}(w_1|w) = \frac{(\alpha+\beta)e^{-(\alpha+\beta)e_1}}{I_{w<0} + e^{-(\alpha+\beta)w}I_{w\geq 0}} \quad \text{for } e_1 > \max\{w, 0\}. \tag{2}$$

Definition 17
Let $Y \sim \mathcal{SL}(\mu, \sigma, \alpha, \beta)$. Then the distribution of $S = \exp(Y)$ is the double Pareto lognormal distribution and in particular, the mean of S is given by

$$E[S] = \frac{\alpha\beta}{(\alpha-1)(\beta+1)} e^{\mu+\frac{\sigma^2}{2}}$$

for $\alpha > 1$.

The density of the $dPlN$ distribution can be easily derived from the skewed Laplace density formula.

Bayesian inference for the $dPlN$ distribution

Reid and Jorgesen (2004) consider classical inference for this model using the EM algorithm. Bayesian inference is examined by Ramírez et al (2008b).

Suppose that we have standard prior distributions:

$$\begin{aligned}
\mu|\sigma^2 &\sim \mathcal{N}\left(m, \frac{\sigma^2}{k}\right) \\
\frac{1}{\sigma^2} &\sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right) \\
\alpha &\sim \mathcal{G}\left(c_\alpha, d_\alpha\right) \\
\alpha &\sim \mathcal{G}\left(c_\beta, d_\beta\right)
\end{aligned}$$

Now, if we observe a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from the $dPlN$ distribution, we can transform the data to $\mathbf{y} = (y_1, \ldots, y_n)$ where $y_i = \log x_i$ and $Y|\mu, \sigma, \alpha, \beta \sim \mathcal{SL}(\mu, \sigma, \alpha, \beta)$.

Clearly, the integration constant of the posterior density $p(\mu, \sigma, \alpha, \beta|\mathbf{y})$ and the marginal densities, $p(\mu|\mathbf{y}), \ldots, p(\beta|\mathbf{y})$ cannot be evaluated analytically. However it is possible to set up a Gibbs sampling scheme by introducing latent variables.

For $i = 1, \ldots, n$, we can define z_i, w_i such that $y_i = z_i + w_i$ and $Z_i|y_i, \mu, \sigma^2$ is generated from the mixture of truncated normal distributions as in Equation 1. Also, we can define w_{i1}, w_{i2} where $w_i = w_{i1} + w_{i2}$ and $W_{i1}|w_i, \alpha, \beta$ has a truncated exponential distribution as in Equation 2.

Conditional on the model parameters, both these distributions are easy to sample.

Conditional on $\mathbf{z} = (z_1, \ldots, z_n)$, then inference for μ, σ^2 is conjugate so that

$$\begin{aligned} \mu|\mathbf{z}, \sigma^2 &\sim \mathcal{N}\left(\frac{km + n\bar{z}}{k+n}, \frac{\sigma^2}{k+n}\right) \\ \frac{1}{\sigma^2}|\mathbf{z} &\sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b + (n-1)s_z^2 + \frac{kn}{k+n}(m-\bar{z})^2}{2}\right). \end{aligned}$$

Conditional on $\mathbf{w}_1 = (w_{11}, \ldots, w_{n1})$ then

$$\alpha|\mathbf{w}_1 \sim \mathcal{G}\left(c_\alpha + n, d_\alpha + n\bar{w_1}.\right)$$

and, conditional on $\mathbf{w}_2 = (w_{21}, \ldots, w_{2n})$, then

$$\beta|\mathbf{w}_2 \sim \mathcal{G}\left(c_\beta + n, d_\beta + n\bar{w_2}.\right).$$

Thus, we can set up a Gibbs sampling algorithm:

Gibbs sampler

1. $t = 0$. Set initial values $\mu^{(0)}, \sigma^{(0)}, \alpha^0, \beta^{(0)}$.
2. For $i = 1, \ldots, n$
   (a) Generate $z_i^{(t)}$ from $f_Z(z|y_i, \mu^{(t-1)}, \sigma^{(t-1)}, \alpha^{t-1}, \beta^{(t-1)})$.
   (b) Set $w_i^{(t)} = y_i - z_i^{(t)}$
   (c) Generate $w_{i1}^{(t)}$ from $f_{W_1}(w_1|w_i^{(t)}, \alpha, \beta)$
   (d) Set $w_{2i}^{(t)} = w_i^{(t)} + w_{1i}^{(t)}$
3. Generate $\mu^{(t)}|\sigma^{(t-1)}, \mathbf{z}^{(t)}$ from $f(\mu|\sigma^{(t-1)}, \mathbf{z}^{(t)})$.
4. Generate $\sigma^{(t)}$ from $f(\sigma|\mathbf{z}^{(t)})$.
5. Generate $\alpha^{(t)}$ from $f(\alpha|\mathbf{w}_1^{(t)})$.
6. Generate $\beta^{(t)}$ from $f(\beta|\mathbf{w}_2^{(t)})$.
7. $t = t + 1$. Go to 2.

Problems

What priors should be used?

The natural choice would be to use the standard, improper priors $p(\mu, \tau) \propto \frac{1}{\tau}$, where $\tau = 1/\sigma^2$, $p(\alpha) \propto \frac{1}{\alpha}$ and $p(\beta) \propto \frac{1}{\beta}$. However, in this case, it is easy to show that the posterior distribution is improper, see e.g. Ramírez et al (2008b). In practice we use small but proper values of all parameters.

What initial values should we use?

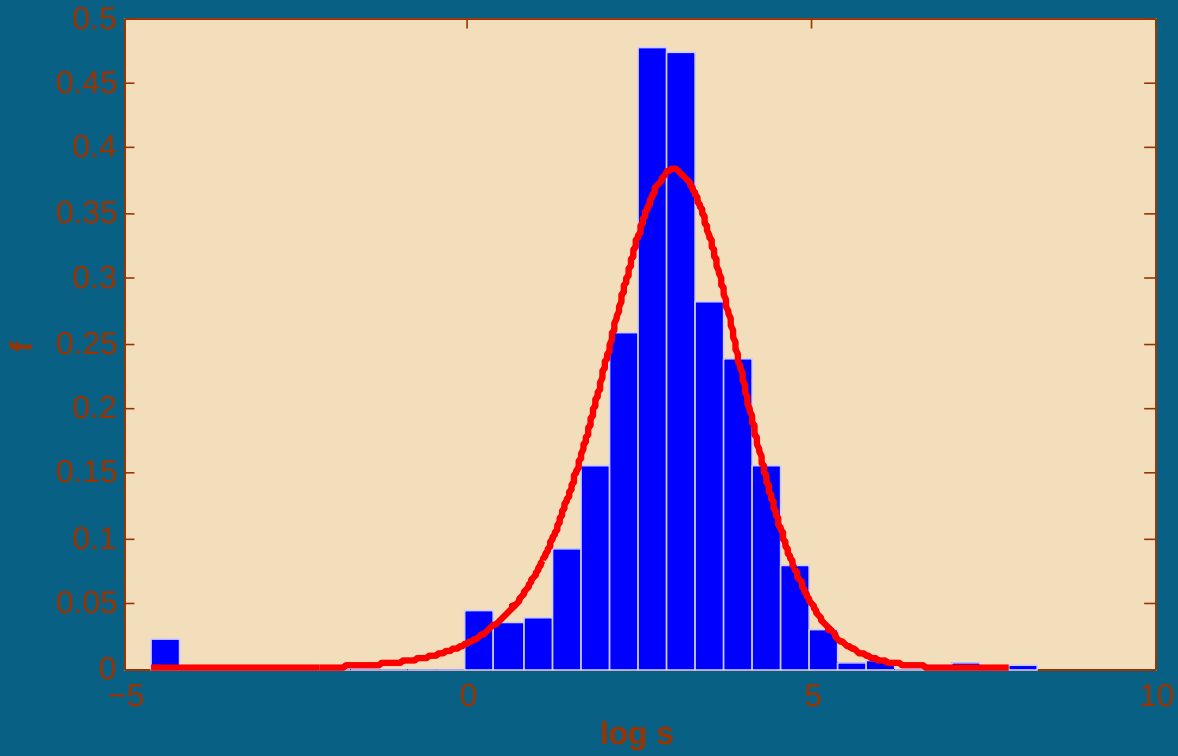
A reasonable choice is to use the maximum likelihood estimates (assuming these exist).

High autocorrelation

We are generating a lot of latent variables here. This leads to high autocorrelation. It is useful to thin the sampled data. We take every hundredth value generated.

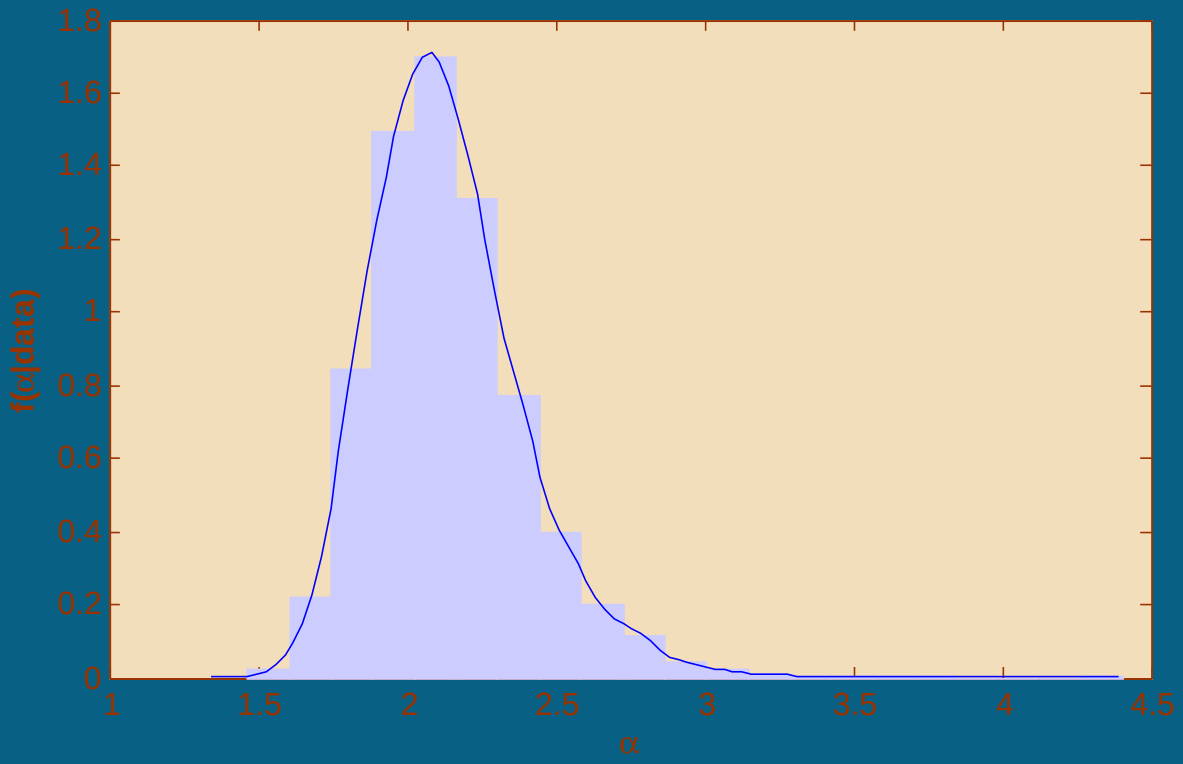
Fitted histogram for the hospital data

The diagram shows the predictive distribution of the logged hospital occupancy times. The fit seems fairly reasonable.



Are the data long tailed?

Recall that the mean of the $dPlN$ distributions only exists if $\alpha > 1$.



The probability that $\alpha < 1$ is virtually zero.

Characteristics of the hospital queueing system

We assume that the hospital has a finite number of beds, c. Patients arrive at the hospital according to a Poisson process with rate λ and are given a bed if one is available and are otherwise lost to the system.

The number of patients in the hospital system can be modeled as a $M/G/c$ Erlang loss system, that is a $M/G/c/c$ system with no queueing, see e.g. Jagerman (1974).

For an Erlang loss system, then the offered load, θ, is defined to be the expected number of arrivals in a service time, that is

$$\theta = \lambda E[S|\mu, \sigma, \alpha, \beta].$$

The equilibrium distribution of the number of occupied beds is given by

$$P(N_b = n|\theta) = \frac{\theta^n/n!}{\sum_{j=0}^{c} \theta^j/j!}$$

and therefore, the blocking probability or probability that an arriving patient is turned away is

$$B(c, \theta) = P(N_b = c|\theta) = \frac{\theta^c/c!}{\sum_{j=0}^{c} \theta^j/j!}$$

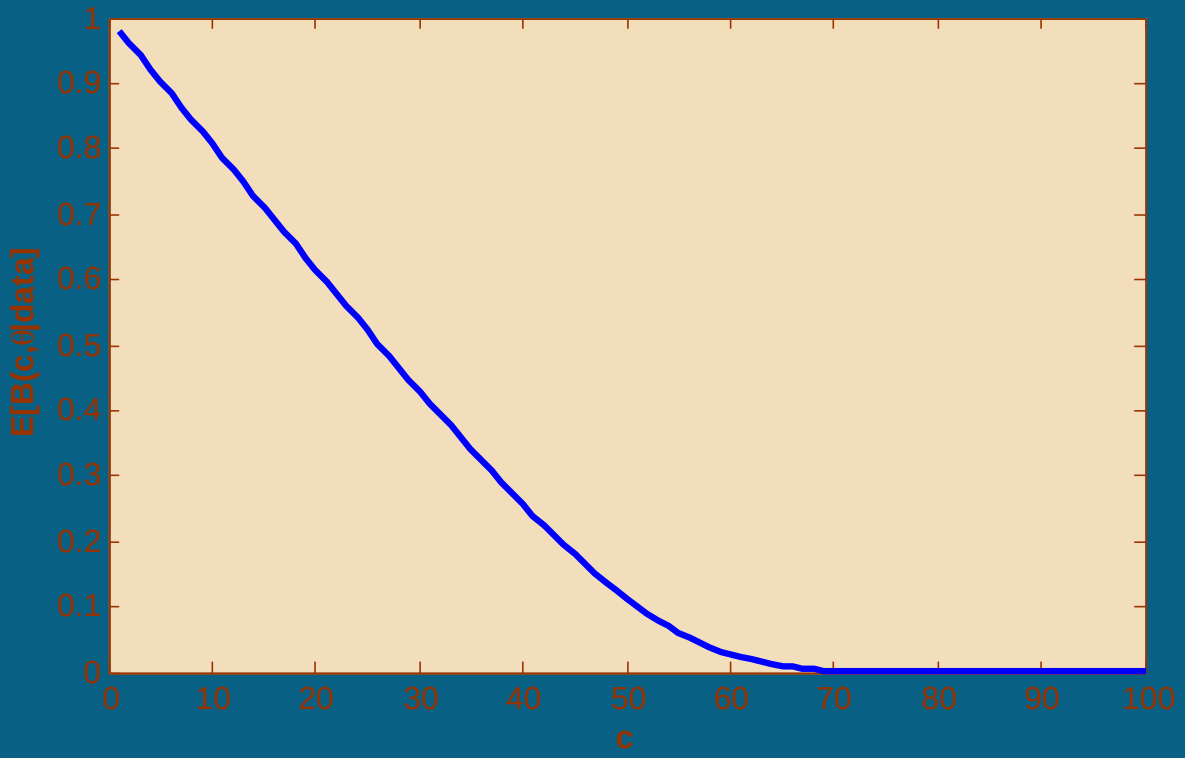
and the expected number of occupied beds is

$$E[N_b|\theta] = \theta\left(1 - B(c, \theta)\right).$$

Given λ, c and a Monte Carlo sample of service time parameters, the above quantities can be estimated by Rao Blackwellization.

Results for the hospital data

Following Ausín et al (2003), we shall suppose that the arrival rate is $\lambda = 1.5$. The diagram shows the blocking probabilities for different numbers of beds.



Optimizing the number of beds

Assume that the hospital accrues different costs for the total numbers of occupied and unoccupied beds and the number of patients that are turned away. The hospital gains profits for those patients treated. Assume then that we have c beds when we shall suppose:

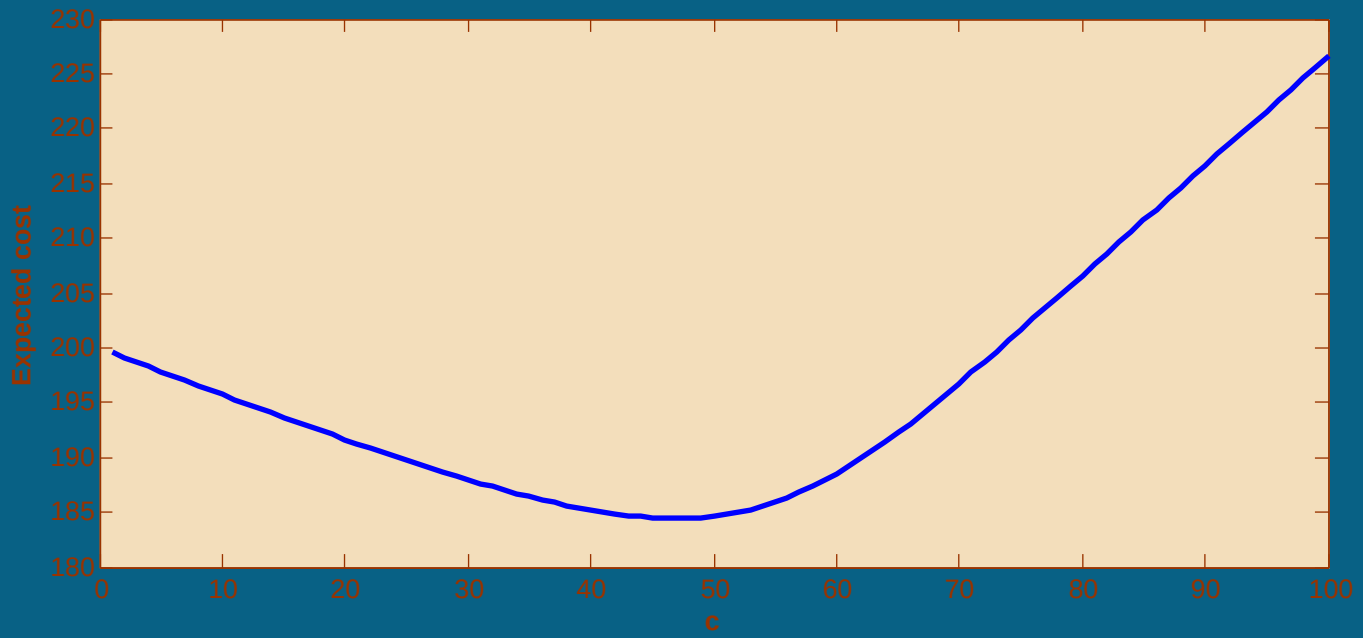
- Cost per occupied bed per time unit is r_b so that the expected cost per time unit due to occupation of beds is $r_b E[N_b|\theta]$.
- Cost r_e per time unit for every empty bed so the expected cost per time unit due to empty beds is $r_e(c - E[N_b|\theta])$.
- Cost per patient turned away per time unit is r_l when the expected cost per time unit is $r_l B(c, \theta)$.

This leads to an expected loss per time unit

$$\begin{aligned} L(c|\lambda,\theta) &= r_b E[N_b|\theta] + r_e(c - E[N_b|\theta]) + r_l B(c,\theta) \\ &= (r_b - r_e)\theta + r_e c + \{(r_e - r_b)\theta + r_l \lambda B(c,\theta)\} \end{aligned}$$

Following Ausín et al (2003), we shall assume that $r_e = 1$, $r_l = 200$ and here we suppose that $r_b = 3$.

Then, we can calculate the number of beds which minimize the expected loss. This is the optimal number of beds from a Bayesian viewpoint.



The optimal number of beds given this loss function is 47 The results are slightly different to those in Ausín et al (2003) who found an optimal number of $c = 58$ with a similar loss function and an alternative service model.

Application IV: Weibull mixture models for heterogeneous survival data

The Weibull distribution is one of the most popular parametric models for survival and reliability. The density function and survival function of a Weibull distributed variable, $X \sim \mathcal{W}(\theta, a)$, are:

$$\begin{aligned} f_W(x|\theta, a) &= \theta a x^{a-1} e^{-\theta x^a} \\ \bar{F}(x|\theta, a) &= e^{-\theta x^a} \end{aligned}$$

The likelihood function

Consider two possible cases:

i We observe n complete lifetimes $\mathbf{x} = x_1, \ldots, x_n$

ii We observe n complete lifetimes as earlier and m truncated lifetimes $x_{m+1}, \ldots, x_{m+n}$ where it is supposed that the subjects are still living at these times.

In case `i.` the likelihood function is

$$l(\theta, a|\mathbf{x}) \propto \theta^n a^n \prod_{i=1}^{n} x_i^a e^{-\theta \sum_{i=1}^{n} x_i^a}$$

and in case `ii`, we have

$$l(\theta, a|\mathbf{x}) \propto \theta^n a^n \prod_{i=1}^{n} x_i^a e^{-\theta \sum_{i=1}^{m+n} x_i^a}$$

Prior and posterior distributions

Suppose that we set the following prior distributions

$$\begin{aligned} \theta &\sim \mathcal{G}(\alpha_\theta, \beta_\theta) \\ a &\sim \mathcal{G}(\alpha_a, \beta_a) \end{aligned}$$

then the conditional posterior distributions are

		Case 1	Case 2
$\theta \vert a, \mathbf{x}$	$\sim$	$\mathcal{G}\left(\alpha_\theta + n, \beta_\theta + \sum_{i=1}^{n} x_i^a\right)$	$\mathcal{G}\left(\alpha_\theta + n, \beta_\theta + \sum_{i=1}^{m+n} x_i^a\right)$
$f(a \vert \theta, \mathbf{x})$	$\propto$	$a^{\alpha_a+n} \prod_{i=1}^{n} x_i^a e^{-\left(\beta_a a + \theta \sum_{i=1}^{n} x_i^a\right)}$	$a^{\alpha_a+n} \prod_{i=1}^{n} x_i^a e^{-\left(\beta_a a + \theta \sum_{i=1}^{m+n} x_i^a\right)}$

Gibbs Sampling

It is easy to set up a Gibbs sampler as follows:

1) $t = 0$. Set initial value $a^{(0)}$.

2) Sample $\theta^{(t+1)}$ from $f\left(\theta|\mathbf{x}, a^{(t)}\right)$

3) Sample $a^{(t+1)}$ from $f\left(a|\mathbf{x}, \theta^{(t+1)}\right)$.

4) $t = t + 1$ Go to 2.

Clearly, step 2 is straightforward. For step 3, Tsionas (2002) uses a Metropolis Hastings step and Marín et al (2005) consider a slice sampler:

A slice sampler for sampling from $f(a|\mathbf{x}, \theta)$

2a) Simulate a uniform random variable; $u \sim \mathcal{U}\left[0, g\left(a^{(t)}|\mathbf{x}, \theta^{(t)}\right)\right]$ where g is the density formula on the previous page

2b) Simulate $a^{(t+1)}$ from a uniform distribution with support $S(u) = \{a : g(a) \geq u\}$.

In practice, the only difficulty with this algorithm is in evaluating the support $S(u)$, although as indicated in Neal (2003), this is straightforward to do by simply sampling from a uniform distribution over a slightly larger space and then checking that the constraint in 2b) is verified.

Mixtures of Weibull distributions

When the data are heterogeneous, it is more appropriate to consider a mixture model

$$f(x|k, \mathbf{w}, \boldsymbol{\theta}, \mathbf{a}) = \sum_{j=1}^{k} w_j f_W(x|\theta_j, a_j).$$

In this case, a natural prior for the weights is $\mathbf{w} \sim \mathcal{D}(\underbrace{c, \ldots, c}_{k})$ and we can use gamma priors for the remaining parameters as earlier.

However, given sample data, e.g. of type 1, then the likelihood becomes

$$l(\mathbf{w}, \boldsymbol{\theta}, \mathbf{a}|\mathbf{x}) \propto \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f_W(x_i|\theta_j, a_j)$$

which contains k^n terms and for n relatively large, is intractable.

Simplifying the likelihood with latent variables

Let Z be a random variable such that $P(Z = z|k, \mathbf{w}) = w_z$. Then, if X comes from the mixture of Weibulls model, we can write

$$X|Z = z \sim \mathcal{W}(\theta_j, a_j).$$

Suppose now that for each observed datum, we know the values $\mathbf{z} = z_1, \ldots, z_n$. Then, the likelihood function simplifies to

$$l(\mathbf{w}, \boldsymbol{\theta}, \mathbf{a}, \mathbf{z}|\mathbf{x}) \propto \prod_{i=1}^{n} f_W(x_i|\theta_{z_i}, a_{z_i}).$$

Posterior distributions

It is now easy to show that

$$\begin{aligned} P(Z_i = z|\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}, \mathbf{a}) &\propto w_z f_W(x_i|\theta_z, a_z) \\ \mathbf{w}|\mathbf{x}, \boldsymbol{\theta}, \mathbf{a}, \mathbf{z} &\sim \mathcal{D}(c + n_1, c + n_2, \ldots, c + n_z) \end{aligned}$$

where n_j are the number of data allocated to element j of the mixture.

The conditional posterior distributions for each θ_j and a_j are as earlier but only considering the sample data assigned to element j of the mixture.

Gibbs sampler

1) $t = 0$. Set initial values $\mathbf{w}^{(0)}, \boldsymbol{\theta}^{(0)}, \mathbf{a}^{(0)}$.

2) For $i = 1, \ldots, n$ sample $z_i^{(t+1)}$ from $P\left(z_i | \mathbf{x}, \mathbf{w}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{a}^{(t)}\right)$.

3) Sample $\mathbf{w}^{(t+1)}$ from $f\left(\mathbf{w} | \mathbf{x}, \mathbf{z}^{(t+1)}, \boldsymbol{\theta}^{(t)}, \mathbf{a}^{(t)}\right)$.

4) For $j = 1, \ldots, k$, sample $\theta_j^{(t+1)}$ from $f\left(\theta_j | \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \mathbf{a}^{(t)}\right)$

5) For $j = 1, \ldots, k$, sample $a_j^{(t+1)}$ from $f\left(a | \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}\right)$.

6) $t = t + 1$ Go to 2.

Inference when k is unknown

In this case, we define a prior distribution for k, e.g. a truncated Poisson. Now, the previous algorithm can be thought of as giving inference for the model parameters conditional on k. We need to incorporate a step which allows for the possibility of changing k and these parameters. There are two possibilities: reversible jump (Richardson and Green 1997) or birth death MCMC (Stephens 2000). Here we use a birth death MCMC approach.

Simulated example

Sample of size 150 with 10% censoring simulated from a mixture of 3 Weibull distributions with weights $\mathbf{w} = (0.6, 0.3, 0.1)$ and parameters $\boldsymbol{\theta} = (0.1, 0.3, 0.5)$ and $\mathbf{a} = (0.5, 1, 2)$.

MCMC algorithm ran for 60000 iterations with 10000 to burn in.

The first graphs illustrate the convergence of the algorithm and the following graph shows the posterior distribution of k. The final graph shows a Kaplan Meier estimate of the survival curve as well as the fitted and true curves

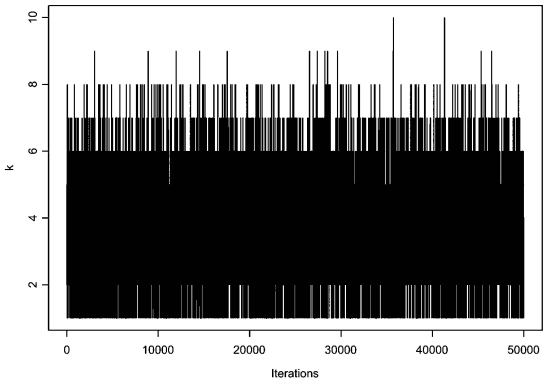


Figure 1. Plot of mixture size k versus iteration of the MCMC algorithm.

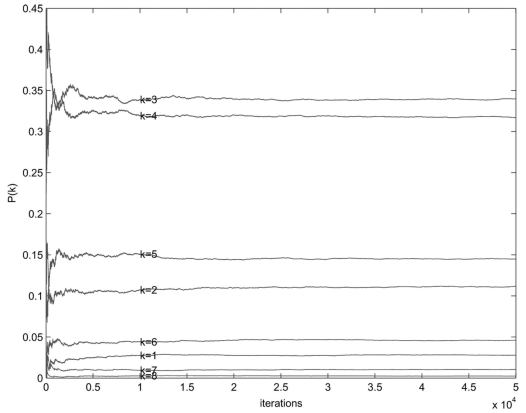


Figure 2. Plot of estimated $P(k \mid \text{data})$ versus iteration of the MCMC algorithm.

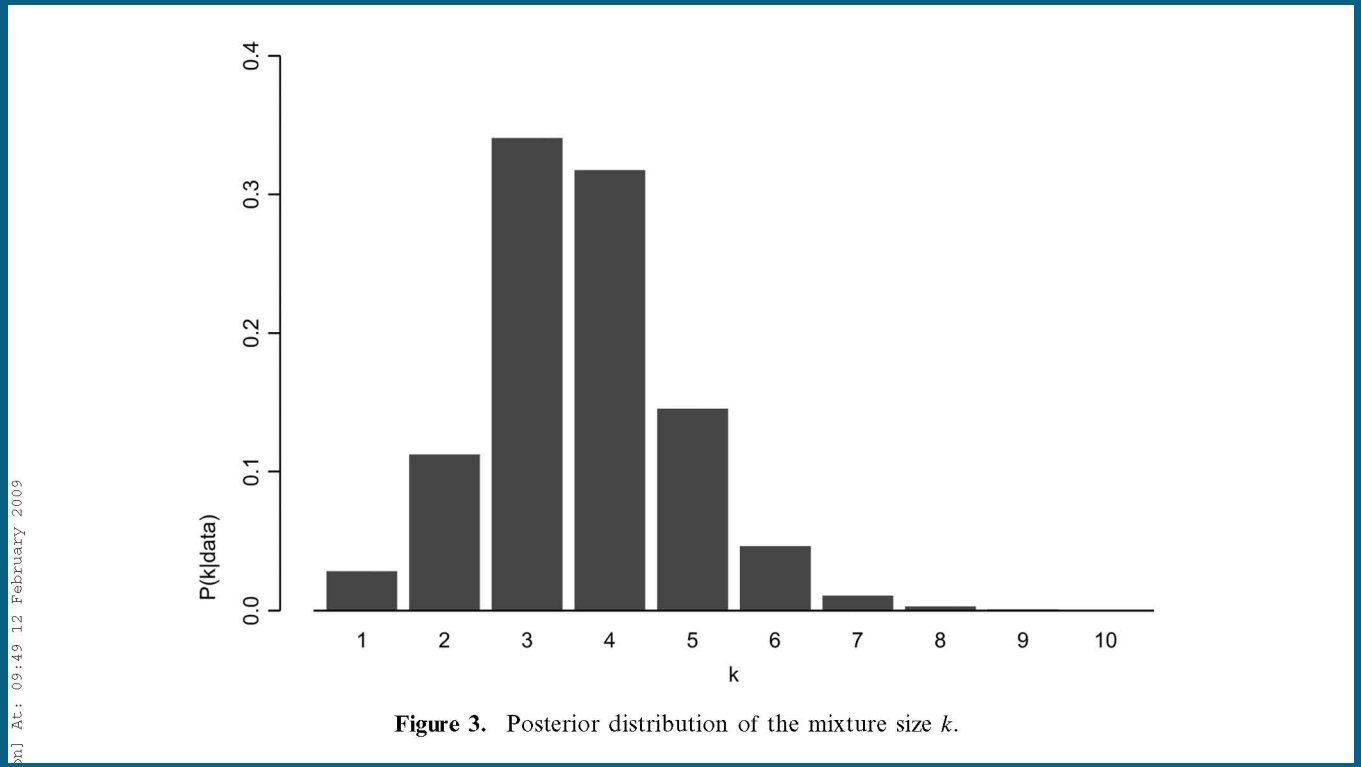


Figure 3. Posterior distribution of the mixture size k.

There is a high posterior probability of 3 components.

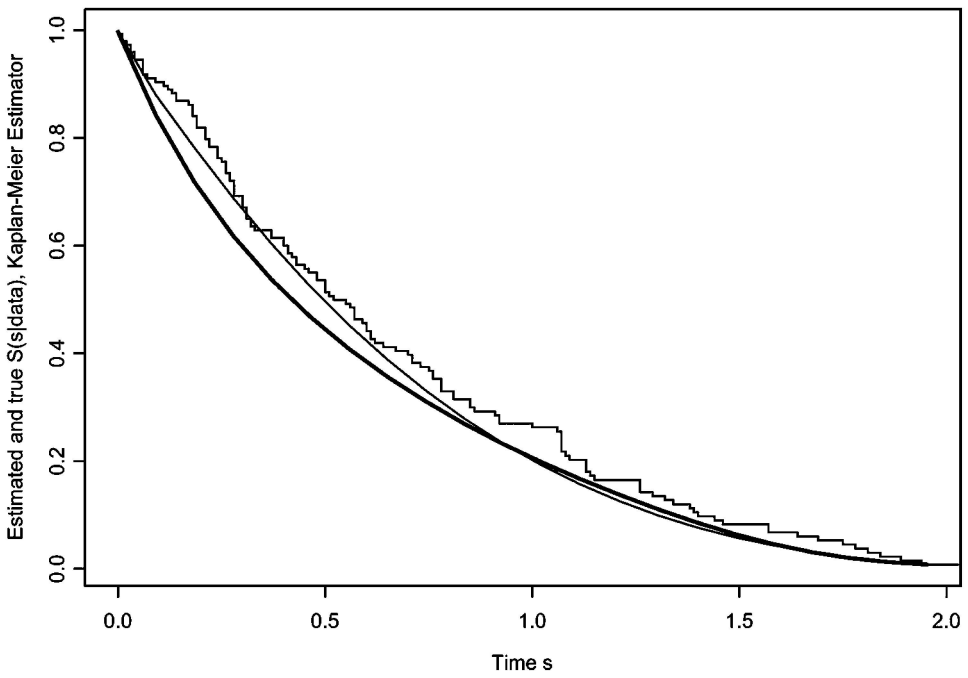


Figure 4. Expected survivor function (thin black line), true survivor function (thick black line) and Kaplan–Meier estimate (polygonal line).

The survival curve is better estimated by the predicted curve than by the KM estimate.

Real data example

Here we analyze data from 87 persons with lupus nephritis, see Abrahamowicz et al. (1996). These patients were studied over a 15-year time period, during which 35 deaths were recorded. In the original article, covariate information was used to study the effects of disease duration prior to diagnosis on the risk of mortality of patients, via a time dependent hazard rate model and suggest that the usual proportional hazards model fits the data reasonably well.

Here, we do not use the covariate information and use the Weibull mixture model to represent the possible inhomogeneity of the lifetime data.

The first graph shows a KM estimate and the fitted survival curve and the second graph shows the estimated hazard curve.

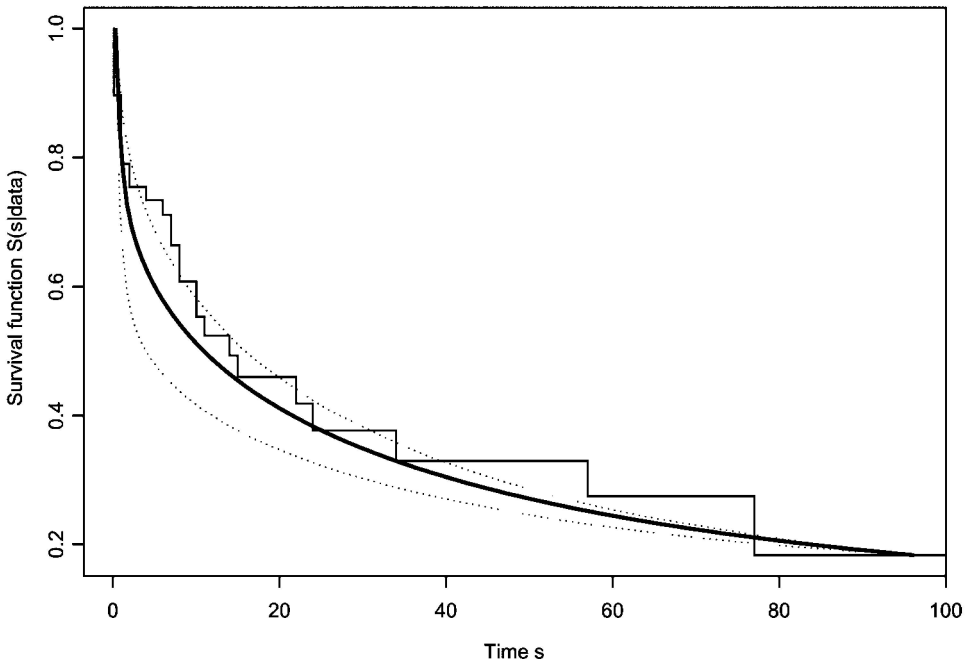


Figure 5. Fitted survival curve and Kaplan–Meier estimator for the Lupus data.

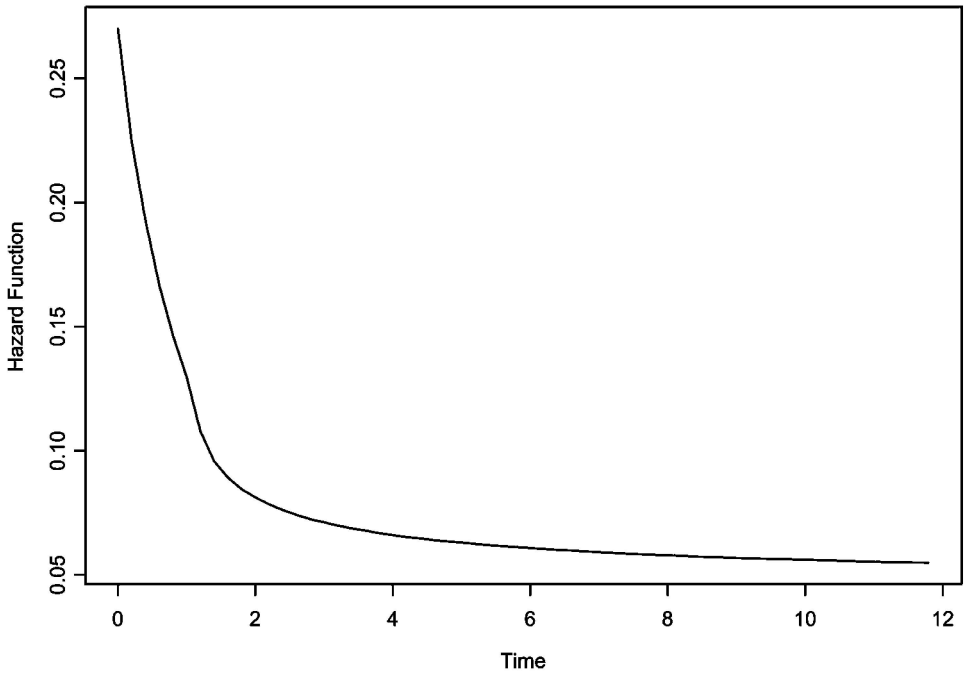


Figure 6. Expected hazard function for the Lupus data.

Finally, in Fig. 6, we illustrate the expected population hazard function for this data set. The expected hazard falls quite rapidly towards 0.05 but then decays very slowly.

These results would seem to suggest that patient death is more probable in the early stages of treatment, but that if a patient survives this phase, then they have a reasonable chance of longer term survival.

References

Abrahamowicz, M., Mackenzie, T., Esdaile, J. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, **91**, 1432-1439.

Ausín, M.C. (2007). An introduction to quadrature and other numerical integration techniques. In *Encyclopedia of Statistics in Quality and Reliability*, Wiley.

Bayesian modelling of hospital bed occupancy times using a mixed generalized Erlang distribution. In Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.), *Bayesian Statistics 7*, Oxford: University Press, 443–452.

Doucet, A., Godsill, S. and Andrieu, C. (2000). On Sequential Monte Carlo Methods for Bayesian Filtering. *Statistics and Computing*, **10**, 197–208. http://www.springerlink.com/content/q6452k2x37357l3r/.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Geweke, J. (1991). Efficient simulation from the Multivariate Normal and Student t-distribution subject to linear constrains. In *Computing Sciences and Statistics (Proc. 23rd Symp. Interface)*, American Statistical Association, 571–577.

Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4*, (eds. Bernardo, J., Berger, J., Dawid, A.P., and Smith, A.F.M.) Oxford University Press.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Gilks, W.R., Best, N.G. and Tan, K.K.C. (1995). Adaptive rejection Metropolis sampling. *Applied Statistics*, **44**, 455–472.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, **57**, 97–109.

Jagerman, D.L. (1974). Some properties of the Erlang loss function, *Bell System Technical Journal*, **53**, 525–551.

Marín, J.M., Rodríguez-Bernal, M.T. and Wiper, M.P. (2005). Using Weibull Mixture Distributions to Model Heterogeneous Survival Data. *Communications in Statistics - Simulation and Computation*, **34**, 673–684.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Neal, R. (2003). Slice sampling (with discussion). *Annals of Statistics*, **31**, 705–767.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms, **9**, 223-252.

Ramirez, P., Lillo, R.E. and Wiper, M.P. (2008a). Bayesian Analysis of a Queueing

System with a Long-Tailed Arrival Process. *Communications in Statistics: Simulation and Computation*, **37**, 697–712.

Ramírez, P., Lillo, R.E., Wiper, M.P. and Wilson, S.P. (2008b). Inference for double Pareto lognormal queues with applications. *Working papers in Statistics and Econometrics*, **08-02**, Universidad Carlos III de Madrid.

Reed, W.J. and Jorgensen, M. (2004). The Double Pareto-Lognormal Distribution - A New Parametric Model for Size Distributions. *Communications in Statistics - Theory and Methods*, **33**, 1733–175.

Robert, C.P. (1995). Simulation of truncated normal random variables. *Statistics and Computing*, **5**, 121–125.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Berlin: Springer.

Roberts, G., Gelman, A. and Gilks, W.R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, **7**, 110–120.

Rubin, D.B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *Journal of the American Statistical Association*, **82**, 543–546.

Tsionas, E. G. (2002). Bayesian analysis of finite mixtures of Weibull distributions. *Communications in Statistics: Theory and Methods*, **31**, 37-48.

Wiper, M.P. (2007). Introduction to Markov chain Monte Carlo simulation. In

Encyclopedia of Statistics in Quality and Reliability, Wiley, 1014-1020.

7. Estimation and hypothesis testing

Objective

In this chapter, we show how the election of estimators can be represented as a decision problem. Secondly, we consider the problem of hypothesis testing from a Bayesian viewpoint and illustrate the similarities and differences between Bayesian and classical procedures.

Recommended reading

- Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Point estimation

Assume that θ is univariate. For Bayesians, the problem of estimation of θ is a decision problem. Associated with an estimator, T, of θ, there is a loss function $L(T, \theta)$ which reflects the difference between the value of T and θ. Then, for an expert, E, with distribution $p(\theta)$, the Bayes estimator of θ is that which minimizes the expected loss

$$E[L(T, \theta)] = \int L(T, \theta) p(\theta) \, d\theta.$$

Definition 18
The *Bayes estimator*, B, of θ is such that

$$E[L(B, \theta)] \leq E[L(T, \theta)] \quad \text{for any } T \neq B.$$

Bayes estimators for some given loss functions

The mean, median and mode of E's distribution for θ can be justified as estimators given certain specific loss functions.

Theorem 29
Suppose that E's density of θ is $p(\theta)$. Then

1. If $L(T,\theta) = (T-\theta)^2$ then B is E's mean, $B = E[\theta]$.

2. If $L(T,\theta) = |T-\theta|$, then B is E's median for θ.

3. If Θ is discrete and $L(T,\theta) = \begin{cases} 0 & \text{if } T = \theta \\ 1 & \text{if } T \neq \theta \end{cases}$ then T is E's modal estimator for θ.

Proof

1.

$$\begin{aligned} E[L(T,\theta)] &= \int (T-\theta)^2 p(\theta)\, d\theta \\ &= \int \left(T - E[\theta] + E[\theta] - \theta\right)^2 p(\theta)\, d\theta \\ &= (T - E[\theta])^2 + V[\theta] \end{aligned}$$

which is minimized when $T = E[\theta] = B$ is the Bayes estimator.

2. Exercise.

3. Exercise.

∎

In the case that Θ is continuous, then if we consider the loss function

$$L(T,\theta) = \begin{cases} 0 & \text{if } |T-\theta| < \epsilon \\ 1 & \text{otherwise} \end{cases},$$

then it is easy to see that T is the centre of the modal interval of width ϵ and letting $\epsilon \longrightarrow 0$, T approaches the mode.

Interval estimation

A expert's 95% credible interval for a variable is simply an interval for which the expert, E, has a 95% probability We have previously used such intervals in the earlier chapters.

Definition 19
If $p(\theta)$ is E's density for θ, we say that (a, b) is a $100(1-\alpha)$**% credible interval** for θ if

$$P(a \leq \theta \leq b|\mathbf{x}) = \int_a^b p(\theta)\, d\theta = 1 - \alpha.$$

It is clear that in general, E will have (infinitely) many credible intervals for θ. The shortest possible credible interval is called a *highest posterior density* (HPD) interval.

Definition 20
The $100 \times (1-\alpha)\%$ HPD interval is an interval of form

$$C = \{\theta : f(\theta) \geq c(\alpha)\}$$

where $c(\alpha)$ is the largest number such that $P(C) \geq 1-\alpha$.

Example 53
$X|\mu \sim \mathcal{N}(\mu, 1)$. Let $f(\mu) \propto 1$. Therefore, $\mu|\mathbf{x} \sim \mathcal{N}(\bar{x}, 1/n)$ and some 95% posterior credible intervals are

$$(-\infty, \bar{x} + 1.64/\sqrt{n}) \quad \text{or} \quad (\bar{x} - 1.64/\sqrt{n}, \infty) \quad \text{or} \quad (\bar{x} \pm 1.96/\sqrt{n})$$

which is the HPD interval.

We can generalize the definition of a credible interval to multivariate densities $p(\boldsymbol{\theta})$. In this case, we can define a credible region $\mathbf{C}$:

$$P(\boldsymbol{\theta} \in \mathbf{C}) = 1 - \alpha.$$

Hypothesis testing

Assume now that we wish to test the hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus the alternative $H_1 : \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$. In theory, this is straightforward. Given a sample of data, $\mathbf{x}$, we can calculate the posterior probabilities $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$, and under a suitable loss function, we can decide to accept or reject the null hypothesis H_0.

Example 54
Given the *all or nothing* loss,

$$L(H_0, \theta) = \begin{cases} 0 & \text{if } H_0 \text{ is true} \\ 1 & \text{if } H_1 \text{ is true} \end{cases}$$

we accept H_0 if $P(H_0|\mathbf{x}) > P(H_1|\mathbf{x})$.

For point null and alternative hypotheses or for one tailed tests, then Bayesian and classical solutions are often similar.

Example 55
Let $X|\theta \sim N(\theta, 1)$. We wish to test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$. Given the uniform prior, $p(\theta) \propto 1$, we know that $\theta|\mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{1}{n}\right)$. Therefore,

$$\begin{aligned} P(H_0|\mathbf{x}) &= P(\theta \leq 0|\mathbf{x}) \\ &= \Phi\left(-\sqrt{n}\bar{x}\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf.

This probability is equal to the usual, classical p-value for the test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$.

$$\begin{aligned} P(\bar{X} \geq \bar{x}|H_0) &= P\left(\sqrt{n}\bar{X} \geq \sqrt{n}\bar{x}|H_0\right) \\ &= 1 - \Phi\left(\sqrt{n}\bar{x}\right) = \Phi\left(-\sqrt{n}\bar{x}\right) \end{aligned}$$

The Lindley/Jeffreys paradox

For two-tailed tests, Bayesian and classical results can be very different.

Example 56
Let $X|\theta \sim \mathcal{N}(\theta, 1)$ and suppose that we wish to test the hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta \neq 0$.

Assume first that

$$p_0 = P(H_0) = 0.5 = P(H_1) = p_1$$

with a normal prior distribution,

$$\theta|H_1 \sim \mathcal{N}(0, 1)$$

and suppose that we observe the mean $\bar{x}$ of a sample of n data with likelihood

$$l(\theta|\bar{x}) = \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2\right).$$

Then, we can calculate the posterior probability, $\alpha_0 = P(H_0|\mathbf{x})$.

$$\begin{aligned}
\alpha_0 &= P(H_0|\bar{x}) \\
&= \frac{p_0 l(\theta = 0|\bar{x})}{p_0 l(\theta = 0|\bar{x}) + p_1 l(\theta|\bar{x})} \\
&= \frac{p_0 l(\theta = 0|\mathbf{x})}{p_0 l(\theta = 0|\mathbf{x}) + p_1 \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta} \\
&= \frac{l(\theta = 0|\mathbf{x})}{l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta} \\
&= K \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right)
\end{aligned}$$

for a constant

$$K^{-1} = l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta.$$

We can also evaluate the second term in the denominator.

$$\begin{aligned} l(\theta|\bar{x})p(\theta|H_1) &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right)\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta^2\right) \\ &= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[n(\bar{x}-\theta)^2+\theta^2\right]\right) \\ &= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[(n+1)\left(\theta-\frac{n\bar{x}}{n+1}\right)^2+\frac{n\bar{x}^2}{n+1}\right]\right) \end{aligned}$$

and, integrating, we have,

$$\int l(\theta|\bar{x})p(\theta|H_1)d\theta = \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right).$$

Therefore, we have

$$\begin{aligned} K^{-1} &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right) + \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right) \\ &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right)\left(1 + \frac{1}{\sqrt{n+1}} \exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right) \end{aligned}$$

so the posterior probability of H_0 is

$$\alpha_0 = \left\{1 + \frac{1}{\sqrt{n+1}} \exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right\}^{-1}.$$

Now suppose that $\bar{x} = 2/\sqrt{n} > 1.96/\sqrt{n}$.

Therefore, for the classical test of H_0 against H_1, with a fixed significance level of 95%, the result is significant and we reject the null hypothesis H_0. However, in this case, the posterior probability is

$$\begin{aligned} \alpha_0 &= \left\{1 + \frac{1}{\sqrt{n+1}} \exp\left(-\frac{2n}{n+1}\right)\right\}^{-1} \\ &\to 1 \quad \text{when } n \to \infty. \end{aligned}$$

Using Bayesian methods, a sample that leads us to reject H_0 using a classical test gives a posterior probability of H_0 that approaches 1 as $n \to \infty$. This is called the Lindley / Jeffreys paradox. See Lindley (1957).

Comments

The choice of the prior variance of θ is quite important but the example also illustrates that it is not very sensible, from a classical viewpoint, to use fixed significance levels as n increases. (In practice, smaller values of α are virtually always used when n increases so that the power of the test is maintained).

Also, point null hypotheses do not appear to be very sensible from a Bayesian viewpoint. The Lindley / Jeffreys paradox can be avoided by considering an interval $(-\epsilon, \epsilon)$ and considering a prior distribution over this interval, see e.g. Berger and Delampady (1987). The main practical difficulty then lies in the choice of ϵ.

Bayes factors

Good

The original idea for the Bayes factor stems from Good (1958) who attributes it to Turing. The motivation is to find a more objective measure than simple posterior probability.

Motivation

An important problem in many contexts such as regression modeling is that of selecting a model $\mathcal{M}_i$ from a given class $\mathcal{M} = \{\mathcal{M}_i : \ i = 1, \ldots, k\}$. In this case, given the a prior distribution $P(\cdot)$ over the model space, we have

$$P(\mathcal{M}_i|\mathbf{x}) = \frac{P(\mathcal{M}_i) f(\mathbf{x}|\mathcal{M}_i)}{\sum_{j=1}^{k} P(\mathcal{M}_j) f(\mathbf{x}|\mathcal{M}_j)},$$

and given the all or nothing loss function, we should select the most probable model.

However, the prior model probabilities are strongly influential in this choice and, if the class of possible models $\mathcal{M}$ is large , it is often difficult or impossible to precisely specify the model probabilities $P(\mathcal{M}_i)$. Thus, it is necessary to introduce another concept which is less strongly dependent on the prior information.

Consider two hypotheses (or models) H_0 and H_1 and let $p_0 = P(H_0)$, $p_1 = P(H_1)$ and $\alpha_0 = P(H_0|\mathbf{x})$ and $\alpha_1 = P(H_1|\mathbf{x})$. Then Jeffreys (1961) defines the Bayes factor comparing the two hypotheses as follows.

Definition 21
The *Bayes factor* in favour of H_0 is defined to be

$$B_1^0 = \frac{\alpha_0/\alpha_1}{p_0/p_1} = \frac{\alpha_0 p_1}{\alpha_1 p_0}.$$

The Bayes factor is simply the posterior odds in favour of H_0 divided by the prior odds. It tells us about the changes in our relative beliefs about the two models caused by the data. The Bayes factor is almost an objective measure and partially eliminates the influence of the prior distribution in that it is *independent* of p_0 and p_1.

Proof

$$\begin{aligned}
\alpha_0 &= P(H_0|\mathbf{x}) = \frac{p_0 f(\mathbf{x}|H_0)}{p_0 f(\mathbf{x}|H_0) + p_1 f(\mathbf{x}|H_1)} \\
\alpha_1 &= \frac{p_1 f(\mathbf{x}|H_1)}{p_0 f(\mathbf{x}|H_0) + p_1 f(\mathbf{x}|H_1)} \\
\frac{\alpha_0}{\alpha_1} &= \frac{p_0 f(\mathbf{x}|H_0)}{p_1 f(\mathbf{x}|H_1)} \\
B_1^0 &= \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}
\end{aligned}$$

which is independent of p_0 and p_1. ∎

The Bayes factor is strongly related to the likelihood ratio, often used in classical statistics for model choice. If H_0 and H_1 are point null hypotheses, then the Bayes factor is exactly equal to the likelihood ratio.

Example 57

Suppose that we wish to text $H_0 : \lambda = 6$ versus $H_1 : \lambda = 3$ and that we observe a sample of size n from an exponential density with rate λ,

$$f(x|\lambda) = \lambda e^{-\lambda x}.$$

Then the Bayes factor in favour of H_0 in this case is

$$\begin{aligned} B_1^0 &= \frac{l(\lambda = 6|\mathbf{x})}{l(\lambda = 3|\mathbf{x})} \\ &= \frac{6^n e^{-6n\bar{x}}}{3^n e^{-3n\bar{x}}} \\ &= 2^n e^{-3n\bar{x}}. \end{aligned}$$

For composite hypotheses however, the Bayes factor depends on the prior parameter distributions. For example, in the case that H_0 is composite, then the marginal likelihood is

$$f(\mathbf{x}|H_0) = \int f(\mathbf{x}|\boldsymbol{\theta}, H_0) p(\boldsymbol{\theta}|H_0)\, d\theta$$

which is dependent on the prior parameter distribution, $p(\boldsymbol{\theta}|H_0)$.

Consistency of the Bayes factor

We can see that the Bayes factor always takes values between 0 and infinity. Furthermore, it is obvious that $B_1^0 = \infty$ if $P(H_0|\mathbf{x}) = 1$ and $B_1^0 = 0$ if $P(H_0|\mathbf{x}) = 0$.

Thus, the Bayes factor is consistent so that if H_0 is true, then $B_1^0 \to \infty$ when $n \to \infty$ and if H_1 is true, then $B_1^0 \to 0$ when $n \to \infty$.

Returning to Example 57, it can be seen that if $\lambda = 6$, then when $n \to \infty$, we have

$$B_1^0 \to 2^n e^{-n/2} \to \infty.$$

Bayes factors and scales of evidence

We can interpret the statistic $2 \log B_1^0$ as a Bayesian version of the classical log likelihood statistic. Kass and Raftery (1995) suggest using the following table of values of B_1^0 to represent the evidence against H_1.

$2 \log_{10} B_1^0$	B_1^0	Evidence against H_1.
0 a 2	1 a 3	Hardly worth commenting
2 a 6	3 a 20	Positive
6 a 10	20 a 150	Strong
> 10	> 150	Very strong

Jeffreys (1961) suggests a (similar) alternative scale of evidence.

Relation of the Bayes factor to standard model selection criteria

The Schwarz (1978) criterion or Bayesian information criterion for evaluating a model, $\mathcal{M}$ is

$$BIC = -2 \log l(\hat{\boldsymbol{\theta}} | \mathcal{M}, \mathbf{x}) + d \log n$$

where $\hat{\boldsymbol{\theta}}$ is the MLE and d is the dimension of the parameter space $\boldsymbol{\Theta}$ for $\mathcal{M}$.

Then, it is possible to show that when the sample size, $n \to \infty$, then

$$BIC_0 - BIC_1 \approx -2 \log B_1^0$$

where BIC_i represents the Bayesian information for model i and B_1^0 is the Bayes factor. See Kass and Raftery (1995).

The deviance information criterion (DIC)

The DIC (Spiegelhalter et al 2002) is a Bayesian alternative to the BIC, AIC etc. appropriate for use in hierarchical models, where the Bayes factor is difficult to calculate. For a model $\mathcal{M}$ and sample data, $\mathbf{x}$, the deviance is:

$$D(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta}) = -2 \log l(\boldsymbol{\theta}|\mathcal{M}, \mathbf{x}).$$

The expected deviance, $\bar{D} = E[D|\mathcal{M}, \boldsymbol{\theta}]$, is a measure of lack of fit of the model. The effective number of model parameters is

$$p_D = \bar{D} - D(\mathbf{x}|\mathcal{M}, \bar{\theta})$$

where $\bar{\theta}$ is the posterior mean. Then the deviance information criterion is $DIC = \bar{D} + p_D$.

An advantage of the DIC is that it is easy to calculate when using Gibbs sampling and this criterion is implemented automatically in `Winbugs`.

For more details, see:

`http://en.wikipedia.org/wiki/Deviance_information_criterion`

However, the DIC has some strange properties which make it inappropriate in certain contexts, e.g. it is not guaranteed that p_D is positive. For alternatives, see e.g. Celeux et al (2006).

Calculation of the Bayes factor

In many cases, when non-conjugate prior distributions are used, the marginal likelihoods $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$ needed for the calculation of the Bayes factor cannot be evaluated analytically. In this case, there are various possibilities:

- Use of an alternative criterion, e.g. BIC or DIC.
- Use of the Laplace approximation. See chapter 8.
- Gibbs sampler or MCMC based approximations of the marginal likelihood. See e.g. Gelfand and Dey (1994), Chib (1995).

Chib's method

Suppose that for a given model, the data are $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$. Given data, $\mathbf{x}$, then we wish to estimate the marginal likelihood $f(\mathbf{x}) = \int l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, where $p(\cdot)$ is the prior for $\boldsymbol{\theta}$.

Assume that $p(\boldsymbol{\theta}|\mathbf{x})$ and therefore $f(\mathbf{x})$, cannot be evaluated analytically, but that we can set up a Gibbs sampler in order to simulate a sample from $p(\boldsymbol{\theta}|\mathbf{x})$.

The assume for simplicity that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and that the conditional distributions $p(\boldsymbol{\theta}_1|\mathbf{x}, \boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\mathbf{x}, \boldsymbol{\theta}_1)$ are available. Chib's method for estimating the marginal likelihood proceeds as follows:

First note that via Bayes theorem, for any $\boldsymbol{\theta}$, we have

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\mathbf{x})}.$$

Therefore,

$$\begin{aligned}\log f(\mathbf{x}) &= \log f(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{x}) \\ &= \log l(\boldsymbol{\theta}|\mathbf{x}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}_2|\mathbf{x}, \boldsymbol{\theta}_1) - \log p(\boldsymbol{\theta}_1|\mathbf{x})\end{aligned}$$

Now, assume that we run a Gibbs sampler for T iterations. Then we can fix some high posterior density point $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ such as the estimated posterior mode or posterior mean. Then all of the terms composing the log marginal likelihood can be estimated directly from the Gibbs sampler output, e.g.

$$\log p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}) \approx \frac{1}{T}\sum_{t=1}^{T} \log p\left(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}, \boldsymbol{\theta}_2^{(t)}\right).$$

If $\boldsymbol{\theta}$ is higher (k) dimensional then the algorithm may be operated in the same way but now writing

$$\log p(\tilde{\boldsymbol{\theta}}|\mathbf{x}) = \log p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}) + \log p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1) + \ldots + \log p(\tilde{\boldsymbol{\theta}}_k|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{k-1}).$$

In this case, $p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x})$ can be estimated directly from the Gibbs output as earlier. In order to estimate $p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1)$, we can run the Gibbs sampler for a further T iterations but holding $\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1$ fixed when we have

$$p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1) \approx \frac{1}{N} \sum_{t=1}^{T} p\left(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}\right).$$

In order to estimate $p(\tilde{\boldsymbol{\theta}}_3|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$, the algorithm is run again but with $\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2 = \tilde{\boldsymbol{\theta}}_2$ fixed and so on. Thus, in order to estimate all the terms, we need to run the Gibbs sampler a total of k times.

In general, this algorithm will be efficient if the point $\tilde{\boldsymbol{\theta}}$ is chosen to have sufficiently high mass, although in theory, it will work for any choice of $\tilde{\boldsymbol{\theta}}$. The disadvantage of the algorithm is the extra execution time needed to run the Gibbs sampler various times.

Chib and Jeliazkov (2001,2005) provide extensions of this algorithm to more complex Markov chain Monte Carlo samplers.

Problems and generalizations of the Bayes factor

If we use improper priors for the model parameters, then in general, the Bayes factor is not defined because

$$B_1^0 = \frac{l(H_0|\mathbf{x})}{l(H_1|\mathbf{x})} = \frac{\int p(\boldsymbol{\theta}_0|H_0) f(\mathbf{x}|\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_0}{\int p(\boldsymbol{\theta}_1|H_1) f(\mathbf{x}|\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

depends on the undefined constants of the prior distributions $p(\cdot|H_0)$ and $p(\cdot|H_1)$.

In such cases, we need to modify the Bayes factor. Various possibilities have been considered.

Intrinsic Bayes factors

Berger

This approach was developed by Berger and Perrichi (1996).

The idea is to divide the sample into two parts; $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ where $\mathbf{y}$ is thought of as training data.

Then we can define the partial Bayes factor in favour of model H_0 based on the data $\mathbf{z}$, after observing $\mathbf{y}$ as

$$B(\mathbf{z}|\mathbf{y}) = \frac{\int f(\boldsymbol{\theta}_0|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_i}{\int f(\boldsymbol{\theta}_1|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

which exists whenever the distributions $f(\boldsymbol{\theta}_i|\mathbf{y})$ are proper for $i = 0, 1$, even though the priors are improper.

A problem with the partial Bayes factor is that this depends on the arbitrary partition of $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. One possibility is to define a measure based on averaging over all of the possible sets $\mathbf{y}$ of the least dimension that gives a proper Bayes factor.

One possibility is to define an arithmetic Bayes factor

$$B_A = \frac{1}{L}\sum_{l=1}^{L} B(\mathbf{z}(l)|\mathbf{y}(l)),$$

where the index l runs over all sets $\mathbf{y}(l)$ of minimum dimension. This has the disadvantage in the model selection context that if we define B^i_j as the Bayes factor in favour of model i as against model j, then $B^i_{Aj} \neq 1/B^j_{Ai}$ so that if we wish to use this Bayes factor in this context, then it is necessary to have a predefined ordering of the models under consideration.

An alternative which does not require such an ordering constraint is the geometric Bayes factor

$$B_G = \left\{\prod_{l=1}^{L} B(\mathbf{z}(l)|\mathbf{y}(l))\right\}^{1/L}.$$

Example 58

Recall Example 55. Suppose that we wish to calculate the Bayes factor in favour of H_0 as against H_1. Earlier, we assumed a uniform prior for θ, i.e.

$$p(\theta|H_0) \propto 1 \text{ and } p(\theta|H_1) \propto 1.$$

Obviously, the Bayes factor is not defined but if we observe a single datum x_l, we have

$$\theta|x_l \sim \mathcal{N}(x_l, 1)$$

which is proper, and

$$\begin{aligned} p(\theta|H_0, x_l) &= \frac{p(\theta|x_l)}{P(\theta \leq 0|x_l)} = \frac{p(\theta|x_l)}{\Phi(-x_l)} \\ p(\theta|H_1, x_l) &= \frac{p(\theta|x_l)}{1 - \Phi(-x_l)} \end{aligned}$$

Therefore, defining $\mathbf{z}(l) = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_n)$, we are able to calculate the intrinsic Bayes factor.

$$
\begin{aligned}
B(\mathbf{z}(l)|x_l) &= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right) \, d\theta}{\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \quad \text{so the intrinsic, geometric Bayes factor is} \\
B^0_{G1} &= \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\prod_{l=1}^{n} \frac{1-\Phi(-x_l)}{\Phi(-x_l)}\right)^{\frac{1}{n}}.
\end{aligned}
$$

Fractional Bayes factors

O' Hagan

This approach stems from O' Hagan (1995).

We define $B_F(b) = \frac{g_0(\mathbf{x}|b)}{g_1(\mathbf{x}|b)}$ where

$$g_i(\mathbf{x}|b) = \int \frac{p(\boldsymbol{\theta}_i|H_i) f(\mathbf{x}|\boldsymbol{\theta}_i)}{\int p(\boldsymbol{\theta}_i|H_i) \{f(\mathbf{x}|\boldsymbol{\theta}_i)\}^b \, d\boldsymbol{\theta}_i} \, d\boldsymbol{\theta}_i$$

for $i = 0, 1$, where b may be interpreted as the proportion of the data chosen for the training sample. We might elect the minimum value of b possible although larger values will produce more robust results.

Example 59

In Example 55, let $b = 1/n$. Therefore

$$g_0(\mathbf{x}|b) = \frac{\int_{-\infty}^{0} \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2n}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}$$

$$= \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \Phi\left(-\sqrt{n}\bar{x}\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \Phi\left(-\bar{x}\right)} \quad \text{and similarly,}$$

$$g_1(\mathbf{x}|b) = \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \left(1-\Phi\left(-\sqrt{n}\bar{x}\right)\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \left(1-\Phi\left(-\bar{x}\right)\right)} \quad \text{and therefore}$$

$$B_F\left(\frac{1}{n}\right) = \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\frac{1-\Phi\left(-\bar{x}\right)}{\Phi\left(-\bar{x}\right)}\right).$$

Fractional and intrinsic Bayes factors do not have all of the properties of simple Bayes factors. See O' Hagan (1997). Also, many variants are available in particular for intrinsic Bayes factors.

Also there are a number of alternative approaches for Bayesian model comparison. See e.g. Wassermann (2000)

http://www.stat.cmu.edu/cmu-stats/tr/tr666/tr666.html

Application II continued: choosing between half-normal and half-t models

In Application II, we considered fitting a half-normal model ($\mathcal{M}_0$) to athletes body fat data and we suggested that a half-t distribution ($\mathcal{M}_1$) might be a reasonable alternative.

Analogous to the half-normal model, we write the half-t model as $X = \xi + \frac{1}{\sqrt{\tau}}|T|$ where $T|d \sim \mathcal{T}_d$ is a Student's t random variable. To simplify the inference, we introduce a latent variable θ such that

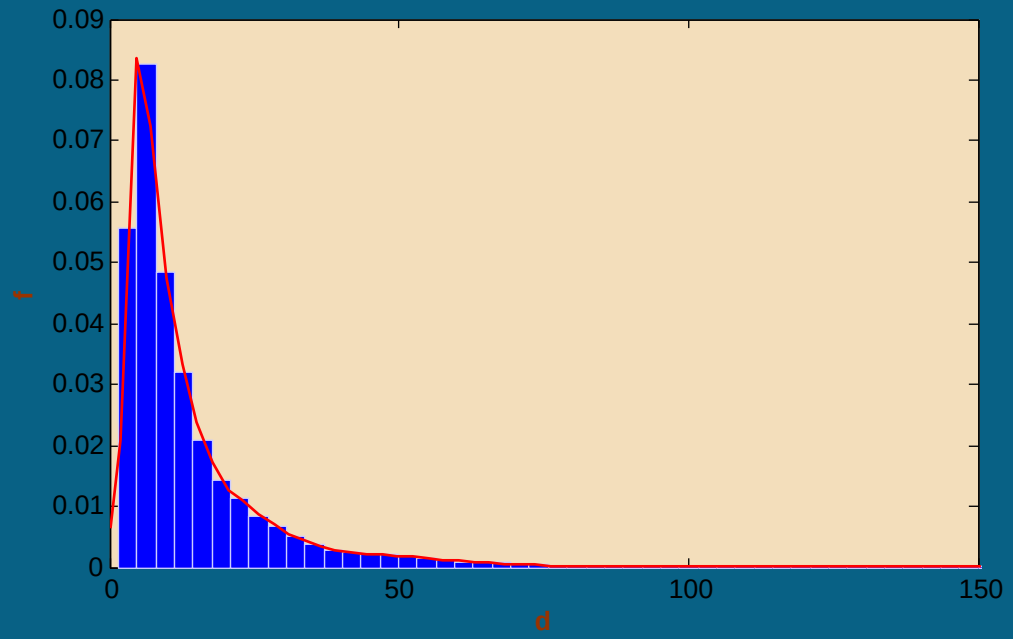
$$\begin{aligned} \theta|d &\sim \mathcal{G}\left(\frac{d}{2}, \frac{d}{2}\right) \quad \text{when} \\ X|\theta, \xi, \tau, \mathcal{M}_1 &\sim \mathcal{HN}\left(\xi, \frac{1}{\tau\theta}\right) \quad \text{has a half-normal distribution.} \end{aligned}$$

Assume that we define improper prior distributions $p(\xi, \tau) \propto \frac{1}{\tau}$ as for the half-normal model and a proper, exponential prior for d, say $d \sim \mathcal{E}(\kappa)$. Then, given a sample, $\mathbf{x}$, the conditional posteriors are:

$$\begin{aligned}
\tau|\mathbf{x}, d, \xi, \boldsymbol{\theta} &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_{i=1}^{n} \theta_i (x_i - \xi_i)^2}{2}\right) \\
\xi|\mathbf{x}, d, \tau, \boldsymbol{\theta} &\sim \mathcal{TN}\left(\frac{\sum_{i=1}^{n} \theta_i x_i}{\sum_{i=1}^{n} \theta_i}, \frac{1}{\tau \sum_{i=1}^{n} \theta_i}\right) \quad \text{truncated so that } \xi < \min\{\mathbf{x}\} \\
\theta_i|\mathbf{x}, d, \xi, \tau &\sim \mathcal{G}\left(\frac{d+1}{2}, \frac{d + \tau(x_i - \xi)^2}{2}\right) \\
p(d|\mathbf{x}, \xi, \tau, \boldsymbol{\theta}) &\propto e^{-\kappa d} \frac{(d/2)^{\frac{nd}{2}}}{\Gamma(d/2)^n} \prod_{i=1}^{n} \theta_i^{\frac{d}{2}} \exp\left(-\frac{d}{2} \sum_{i=1}^{n} \theta_i\right)
\end{aligned}$$

The only density that has a slightly complicated form is that of d and the joint posterior can be sampled using a Gibbs algorithm with a *Metropolis* step for d.

The histogram shows an estimate of the posterior of d when the half-t model was fitted to the athletes data with a prior distribution $d \sim \mathcal{E}(1/20)$.



The distribution is quite long tailed although the mode of d is below 10. We saw earlier that the fit of the half-t model looks better than the half-normal fit, but how can we calculate a Bayes factor?

Calculation of the Bayes factor

We saw earlier that, in general if improper prior distributions are used, then the Bayes factor is undefined. However, in this case, the usual Bayes factor construction:

$$\begin{aligned} B_1^0 &= \frac{f(\mathbf{x}|\mathcal{M}_0)}{f(\mathbf{x}|\mathcal{M}_1)} \\ &= \frac{\int f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)p(\xi,\tau|\mathcal{M}_0)\,d\xi d\tau}{\int f(\mathbf{x}|d,\xi,\tau,\mathcal{M}_1)p(\xi,\tau|\mathcal{M}_1)p(d|\mathcal{M}_1)\,d\xi d\tau, dd} \\ &= \frac{\int_{-\infty}^{\min\{\mathbf{x}\}}\int_0^\infty f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)\frac{1}{\tau}\,d\xi d\tau}{\int_0^\infty\int_{-\infty}^{\min\{\mathbf{x}\}}\int_0^\infty\int f(\mathbf{x}|d,\xi,\tau,\mathcal{M}_1)\frac{1}{\tau}p(d|\mathcal{M}_1)\,d\xi d\tau dd} \end{aligned}$$

can be shown to produce a well-calibrated, intrinsic Bayes factor because the support and improper priors for ξ, τ are the same under $both$ models. See Cano et al (2004).

The numerator of this formula can be calculated explicitly for the half-normal model. Thus:

$$\int_{-\infty}^{\min\{\mathbf{x}\}} \int_{0}^{\infty} f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)\frac{1}{\tau}\,d\xi d\tau = \frac{2\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n}}\left(\frac{2}{\sqrt{(n-1)\pi}s}\right)^{n-1}\Phi_{n-1}\left(\frac{\min\{\mathbf{x}\}-\bar{x}}{s/\sqrt{n}}\right).$$

The denominator can be calculated using Chib's approach. We can write

$$\begin{aligned}\log f(\mathbf{x}|\mathcal{M}_1) &= \log p(\tilde{\xi},\tilde{\tau}) + \log p(\tilde{d}) + \log l(\tilde{\xi},\tilde{\tau},\tilde{d}|\mathbf{x}) \\ &\quad - \log p(\tilde{\xi}|\mathbf{x}) - \log p(\tilde{\tau}|\mathbf{x},\tilde{\xi}) - \log p(\tilde{d}|\mathbf{x},\tilde{\xi},\tilde{\tau})\end{aligned}$$

and the integrating constant of the density in the final term in this expression can be evaluated by using standard, one dimensional, numerical integration.

Using this approach, the Bayes factor in favour of the half-t model is $B_0^1 = 4$ which suggests positive evidence in favour of this model. For more details, see Wiper et al (2008).

References

Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, **2**, 217–352.

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Cano, J.A., Kessler, M. and Moreno, E. (2004). On intrinsic priors for nonnested models. *Test*, **13**, 445–463.

Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006). Deviance Information Criteria for Missing Data Models (with discussion). *Bayesian Analysis*, **1**, 651–706.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis Hastings output. *Journal of the American Statistical Association*, **96**, 270–281.

Chib, S. and Jeliazkov, I. (2005). Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, **59**, 30–44.

Gelfand, A.E., and Dey, D.K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.

Good, I.J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, **53**, 799–813.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.), Oxford: University Press.

Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Lindley, D.V. (1957). A Statistical Paradox. *Biometrika*, **44**, 187-192.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 99–138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, **6**, 101–118.

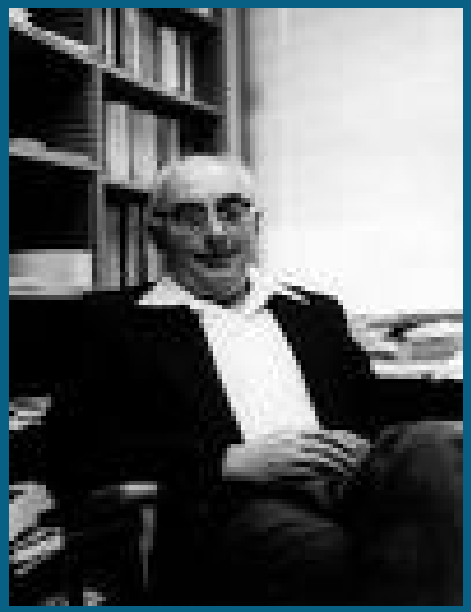
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D.J., Best, N.G., Carlin B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583-640.

Wassermann, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Wiper, M.P., Girón, F.J. and Pewsey, A. (2008). Objective Bayesian inference for the half-normal and half-t distributions. *Communications in Statistics: Theory and Methods*, **37**, 3165–3185.

8. Large samples



Le Cam

Le Cam (1953) was the first to formally demonstrate the asymptotic normality of the posterior distribution.

Objective

Illustrate the limiting properties of Bayesian distributions.

Recommended Reading

- Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory*, Section 5.3.

- Gelman et al (2003), Chapter 4, Sections 4.1 – 4.3 and Appendix B.

If the sample size is very large, it seems obvious that the prior parameter values will have very little influence.

Example 60
$X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we use a conjugate prior distribution. Then, for example:

$$E[\mu|\mathbf{x}] = \frac{cm + n\bar{x}}{c + n} \to \bar{x}$$

when $n \to \infty$.

In fact, we should usually expect that the properties of Bayesian posterior distributions will be similar to those of maximum likelihood estimators in the limit. The following results illustrate this.

Asymptotic results when $\boldsymbol{\Theta}$ is discrete

The following theorem demonstrates that, in the discrete case, as long as the prior probability of the true value, $\boldsymbol{\theta}_t$, is positive, then the posterior probability density of $\boldsymbol{\theta}$ converges to a point mass at $\boldsymbol{\theta}_t$.

Theorem 30
Let $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ where the parameter space $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$ is countable. Suppose that $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}$ is the true value of θ.

Suppose that the prior distribution is $P(\boldsymbol{\theta})$ where $P(\boldsymbol{\theta}_i) > 0 \; \forall \; i$ and we assume that

$$\int f(x|\boldsymbol{\theta}_t) \log \frac{f(x|\boldsymbol{\theta}_t)}{f(x|\boldsymbol{\theta}_i)} \, dx > 0 \quad \forall \; i \neq t. \quad \text{Then}$$

$\lim_{n\to\infty} P(\boldsymbol{\theta}_t|\mathbf{x}) = 1$ and $\lim_{n\to\infty} P(\boldsymbol{\theta}_i|\mathbf{x}) = 0 \; \forall \; i \neq t$.

An interesting extension of this result is that if $\boldsymbol{\theta}_t \notin \boldsymbol{\Theta}$, then the posterior distribution converges to a point mass at the point that gives a parametric model closest (in the sense of Kullback-Liebler distance) to the true model.

Proof Let $\mathbf{x} = (x_1, \ldots, x_n)$.

$$\begin{aligned} P(\boldsymbol{\theta}_i|\mathbf{x}) &= \frac{P(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)}{\sum_i P(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)} \\ &= \frac{P(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)/f(\mathbf{x}|\boldsymbol{\theta}_t)}{\sum_i P(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)/f(\mathbf{x}|\boldsymbol{\theta}_t)} \\ &= \frac{\exp\left(\log P(\boldsymbol{\theta}_i) + S_i\right)}{\sum_i \exp\left(\log P(\boldsymbol{\theta}_i) + S_i\right)} \quad \text{where } S_i = \sum_{j=1}^{n} \log \frac{f(x_j|\boldsymbol{\theta}_i)}{f(x_j|\boldsymbol{\theta}_t)}. \end{aligned}$$

Conditional on $\boldsymbol{\theta}_t$, S_i is the sum of n i.i.d. random quantities and therefore, by the strong law of large numbers, we have

$$\lim_{n\to\infty} \frac{1}{n} S_i = \int f(x|\boldsymbol{\theta}_t) \log \frac{f(x|\boldsymbol{\theta}_i)}{f(x|\boldsymbol{\theta}_t)}\, dx.$$

This quantity is negative if $i \neq t$ and zero if $i = t$. Therefore, when $n \to \infty$, $S_t \to 0$ y $S_i \to -\infty$ if $i \neq t$, which proves the theorem. ∎

The continuous case

The previous arguments cannot be used in the continuous case, as now, the probability at any particular value of $\boldsymbol{\theta}$ is 0. Instead, we now define $\boldsymbol{\theta}_t$ to be the value of $\boldsymbol{\theta}$ that maximizes the Kullback-Liebler information

$$H(\boldsymbol{\theta}) = \int \log \frac{f_t(x)}{f(x|\boldsymbol{\theta})} f_t(x)\, dx$$

of the distribution $f(\cdot|\boldsymbol{\theta})$ with respect to the true distribution of X, say $f_t(\cdot)$. Now we can demonstrate the following theorem.

Theorem 31
If $\boldsymbol{\theta}$ is defined on a compact set and A is a neighbourhood of $\boldsymbol{\theta}_t$ with non-zero prior probability, then

$$P(\boldsymbol{\theta}_t \in A|\mathbf{x}) \to 1 \quad \text{as } n \to \infty.$$

Proof See Gelman et al (2003). ∎

Convergence to normality

Theorem 32
Under certain regularity conditions, the posterior distribution of $\boldsymbol{\theta}$ tends to a normal distribution with mean $\boldsymbol{\theta}_t$ and variance $nJ(\boldsymbol{\theta}_t)^{-1}$ where $J(\boldsymbol{\theta})$ is the Fisher information.

Proof Suppose that θ is univariate and let $\hat{\theta}$ be the posterior mode. Then a Taylor expansion of $\log p(\theta|\mathbf{x})$ around the mode is

$$\log p(\theta|\mathbf{x}) = \log p(\hat{\theta}|\mathbf{x}) + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{x})|_{\theta=\hat{\theta}} + \dots$$

The first term in this expression is constant and the second term is

$$\begin{aligned}
(\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log p(\theta|\mathbf{x})|_{\theta=\hat{\theta}} &= (\theta - \hat{\theta})^2 \frac{d^2}{d\theta^2} \log \left. \frac{p(\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x})} \right|_{\theta=\hat{\theta}} \\
&= (\theta - \hat{\theta})^2 \left(\frac{d^2}{d\theta^2} \log p(\hat{\theta}) + \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)|_{\theta=\hat{\theta}} \right).
\end{aligned}$$

Now the first bracketed term, $\frac{d^2}{d\theta^2} \log p(\hat{\theta})$ is constant, whereas the second term (thinking of the x_i as variables) is the sum of n i.i.d. random variables with negative mean.

As the posterior mode is a consistent estimator, it follows that if $f_t(x) = f(x|\theta_t)$ is the true distribution, then the mean is $-J(\theta_t)$. Otherwise, the mean is $E_{f_t}\left[\frac{d^2}{d\theta^2} \log f(x|\theta)\right]$ evaluated at $\theta = \theta_t$ which is also negative by definition of θ_t.

Therefore, the coefficient of the second term in the Taylor series increases with order n. Similarly, the higher order terms can also be shown to increase no faster than order n.

Letting $n \to \infty$, we thus have that the importance of the higher order terms of the Taylor expansion fades relative to the quadratic term as the mass of the posterior concentrates around θ_t and the normal approximation grows in precision. ∎

Here, we used the mode as a consistent estimator of $\boldsymbol{\theta}_t$. We could equally use the mean or the classical MLE.

Theorem 33
Let $X_i|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ with prior distribution $f(\boldsymbol{\theta})$. Given data $\mathbf{x}$, when $n \to \infty$,

1. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(E[\boldsymbol{\theta}|\mathbf{x}], V[\boldsymbol{\theta}|\mathbf{x}])$, supposing that the mean and variance of $\boldsymbol{\theta}$ exist,

2. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, I(\hat{\boldsymbol{\theta}})^{-1})$ where $\hat{\boldsymbol{\theta}}$ is the mode. $I(\boldsymbol{\theta})$ is *the observed information*

$$I(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(\boldsymbol{\theta}|\mathbf{x})).$$

3. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\hat{\boldsymbol{\theta}}}, I^{\star}(\hat{\hat{\boldsymbol{\theta}}})^{-1})$ where $\hat{\hat{\boldsymbol{\theta}}}$ is the MLE of $\boldsymbol{\theta}$, supposing this exists and

$$I^{\star}(\boldsymbol{\theta}) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(\mathbf{x}|\boldsymbol{\theta}))$$

4. $\boldsymbol{\theta}|\mathbf{x} \approx \mathcal{N}(\hat{\hat{\boldsymbol{\theta}}}, I^{\star\star}(\hat{\hat{\boldsymbol{\theta}}})^{-1})$ where $I^{\star\star}(\boldsymbol{\theta}) = -nE_X\left[\frac{d^2}{d\boldsymbol{\theta}^2} \log(f(X|\boldsymbol{\theta}))\right]$.

Proof See Bernardo and Smith (1994) or Gelman et al (2003). ∎

Approximating a beta posterior distribution

Normally, the first approximation will be better than the second and so on. In many cases, the posterior mean and variance are difficult to evaluate but it is much easier to calculate the mode and observed information.

Example 61
Let $X|\theta \sim \mathcal{BI}(n, \theta)$ and $\theta \sim \mathcal{B}(\alpha, \beta)$. Then, $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + x, \beta + n - x)$. If n is large, we can approximate the posterior distribution of θ. Here we compare the four approximations given earlier.

Firstly, approximating with a normal using the beta mean and variance we have

$$\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{\alpha + x}{\alpha + \beta + n}, \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}\right).$$

Secondly, we can calculate the mode of the beta distribution.

$$\begin{aligned}
\log p(\theta|\mathbf{x}) &= c + (\alpha + x - 1)\log(\theta) + (\beta + n - x - 1)\log(1-\theta) \\
\frac{d}{d\theta} &= \frac{\alpha + x - 1}{\theta} - \frac{\beta + n - x - 1}{1-\theta} \\
\hat{\theta} &= \frac{\alpha + x - 1}{\alpha + \beta + n - 2}
\end{aligned}$$

is the mode. Also:

$$\begin{aligned}
\frac{d^2}{d\theta^2}\log(f(\theta|\mathbf{x})) &= -\frac{\alpha + x - 1}{\theta^2} - \frac{\beta + n - x - 1}{(1-\theta)^2} \\
I(\hat{\theta}) &= \frac{\alpha + x - 1}{\hat{\theta}^2} + \frac{\beta + n - x - 1}{(1-\hat{\theta})^2} \\
&= \frac{(\alpha + \beta + n - 2)^3}{(\alpha + x - 1)(\beta + n - x - 1)} \quad \text{and therefore} \\
\theta|\mathbf{x} &\approx \mathcal{N}\left(\frac{\alpha + x - 1}{\alpha + \beta + n - 2}, \frac{(\alpha + x - 1)(\beta + n - x - 1)}{(\alpha + \beta + n - 2)^3}\right)
\end{aligned}$$

Thirdly, note that the MLE is $\hat{\hat{\theta}} = \frac{x}{n}$ and $I^{\star}(\hat{\hat{\theta}}) = \frac{n^3}{x(n-x)}$. Thus

$$\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{x}{n}, \frac{x(n-x)}{n^3}\right)$$

and finally, note that

$$\begin{aligned}
-\frac{d^2}{d\theta^2}\log(f(X|\theta)) &= \frac{X}{\theta^2} + \frac{n-X}{(1-\theta)^2} \\
I^{\star\star}(\theta) &= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} \\
&= \frac{n}{\theta} + \frac{n}{(1-\theta)} \\
I^{\star\star}(\hat{\hat{\theta}}) &= \frac{n^3}{x(n-x)}
\end{aligned}$$

and we have the same approximation.

When $n >> \alpha + \beta$ the approximations will give similar results but in small samples, the results could be quite different.

Example 62

Suppose that $\alpha = \beta = 2$ and $x = 20$, $n = 30$. We shall use the different approximations and estimate $P(\theta > 0.5|\text{data})$.

We have $\theta|\mathbf{x} \sim \mathcal{B}(22, 12)$ and using Matlab, we can show that $P(\theta > 0.5|\mathbf{x}) = 0.95993$ is the exact probability.

Now, using the first approximation, we have $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{22}{34}, \frac{22\times12}{34^2\times35}\right) = \mathcal{N}(0.64706, 0.006525)$ and we find $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.8206) = 0.9660$.

Approximating using the mode, we have $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{21}{32}, \frac{21\times11}{32^3}\right) = \mathcal{N}(0.65625, 0.00705)$ and thus $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.8610) = 0.9686$.

Using the classical approximations, $\theta|\mathbf{x} \approx \mathcal{N}\left(\frac{20}{30}, \frac{20\times10}{30^3}\right) = \mathcal{N}(0.66667, 0.00741)$ and thus, $P(\theta > 0.5|\mathbf{x}) \approx P(Z > -1.9365) = 0.9735$.

When the theorem cannot be applied

In some situations we cannot apply the results of the theorem. For example

- If $\boldsymbol{\theta}_t$ is a boundary point of $\boldsymbol{\Theta}$,
- If the prior mass density around $\boldsymbol{\theta}_t$ is 0,
- If the posterior density is improper,
- If the model is not identifiable.

Example 63
Suppose that we have the model

$$f(x|\theta_1, \ldots, \theta_k) = w_1 f(x|\theta_1) + \ldots + w_k f(x|\theta_k)$$

i.e. a mixture of k densities from the same family.

Then given the data, the likelihood will be multimodal because the model is not identifiable. Thus, we need to restrict the parameter space Θ in order to identify the model.

One possibility is to order the parameters, $\theta_1 < \ldots < \theta_k$, to identify the model.

See Gelman et al (2003) for more examples.

The Laplace approximation

Tierney and Kadane (1996) introduced this generalization of the normal approximation in order for the problem of estimating posterior moments.

Assume that we wish to estimate

$$E[g(\theta)|\mathbf{x}] = \frac{\int g(\theta) l(\theta|\mathbf{x}) p(\theta)\, d\theta}{\int l(\theta|\mathbf{x}) p(\theta)\, d\theta}$$

where it is supposed that $g(\cdot)$ is non negative.

Then we can write this expectation as

$$E[g(\theta)|\mathbf{x}] = \frac{\int \exp(-nh^*(\theta))\, d\theta}{\int \exp(-nh(\theta))\, d\theta} \quad \text{where}$$

$$\begin{aligned} -nh(\theta) &= \log p(\theta) + \log l(\theta|\mathbf{x}) \\ \text{and} \quad -nh^*(\theta) &= \log g(\theta) + \log p(\theta) + \log l(\theta|\mathbf{x}). \end{aligned}$$

Then, we use the Taylor expansion of h (h^*) about the mode $\hat{\theta}$ ($\hat{\hat{\theta}}$).

$$-h(\hat{\theta}) = \max_{\theta}(-h(\theta)) \qquad -h^*(\hat{\hat{\theta}}) = \max_{\theta}(-h^*(\theta))$$

and retain the quadratic terms. We estimate the denominator by

$$\int \exp(-nh(\theta))\, d\theta \approx \sqrt{2\pi}\sigma n^{-1/2} \exp(-nh(\hat{\theta}))$$

where $\sigma = \left(\frac{d^2}{d\theta^2} h(\theta)|_{\theta=\hat{\theta}}\right)^{-1/2}$ and similarly for the numerator.

This leads to the following estimate:

$$E[g(\theta|\mathbf{x})] \approx \left(\frac{\sigma^*}{\sigma}\right) \frac{g(\hat{\hat{\theta}}) f(\hat{\hat{\theta}}) l(\hat{\hat{\theta}}|\mathbf{x})}{f(\hat{\theta}) l(\hat{\theta}|\mathbf{x})} \quad \text{where}$$

$$\sigma^* = \left(\frac{d^2}{d\theta^2} h^*(\theta)|_{\theta=\hat{\hat{\theta}}}\right)^{-1/2}.$$

Example 64

Return to Example 61. We have $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + x, \beta + n - x)$.

Without loss of generality, suppose that $0 \leq \alpha, \beta < 1$. If not, simply transform, $x \to x + [\alpha]$ and $n \to n + [\alpha] + [\beta]$.

Writing the beta density as above,

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \\ &\propto \exp(-nh(\theta)) \quad \text{where} \\ h(\theta) &= -\frac{1}{n}\left((\alpha+x-1)\log\theta + (\beta+n-x-1)\log(1-\theta)\right) \end{aligned}$$

We can thus show (Exercise) that the Laplace estimate of the posterior mean will be

$$\frac{(\alpha+x)^{\alpha+x+1/2}}{(\alpha+x-1)^{\alpha+x-1/2}} \frac{(\alpha+\beta+n-2)^{\alpha+\beta+n-1/2}}{(\alpha+\beta+n-1)^{\alpha+\beta+n+1/2}}.$$

For example, if $\theta|\mathbf{x}\ \mathcal{B}(8,12)$, setting $\alpha = \beta = 0$ and $n = 20$ the Laplace estimate of the posterior mean is

$$E[\theta|\mathbf{x}] \approx \frac{8^{8.5}}{7^{7.5}} \frac{18^{19.5}}{19^{20.5}} \approx .3994$$

The true value of the mean is $8/20 = 0.4$ and approximating the mean by the mode as in approximation 2 of the theorem, we have $E[\theta|\mathbf{x}] \approx 7/18 = .3889$. The Laplace approximation is somewhat better.

Approximating the Bayes factor

Consider the case of two composite hypotheses H_0 and H_1. The Bayes factor is

$$B = \frac{\int f(\mathbf{x}|\boldsymbol{\theta}_0, H_0) f(\boldsymbol{\theta}_0|H_0)\, d\boldsymbol{\theta}_0}{\int f(\mathbf{x}|\boldsymbol{\theta}_1, H_1) f(\boldsymbol{\theta}_1|H_1)\, d\boldsymbol{\theta}_1}$$

and both numerator and denominator are positive functions. Therefore we can apply the Laplace approximation. See Kass and Raftery (1995) for details.

Properties and problems with the Laplace approximation

- The Laplace approximation is $O(1/n^2)$.
- If $\Theta \neq R$, then the model can be reparameterized in order to improve the approximation.
- The Laplace approximation can be extended to the multivariate situation.
- In order to implement the Laplace approximation, we need to be able to calculate the MLE of θ.

References

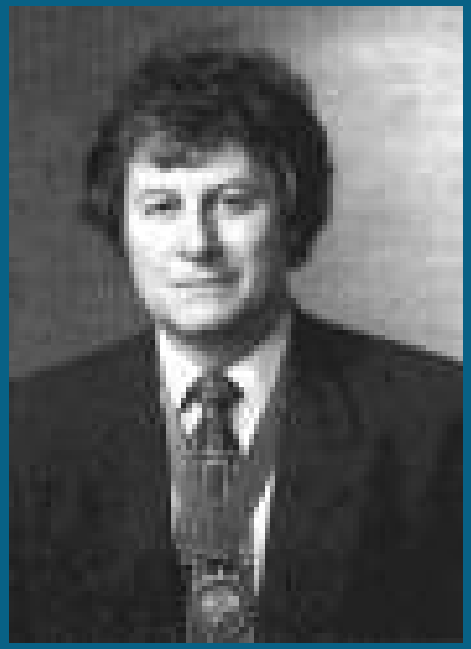
Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (2003). *Bayesian Data Analysis* (2'nd ed.), Chapman and Hall.

Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Le Cam, L. (1953). On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates. *University of California Publications in Statistics*, **1**, 277–328.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

9. Linear models and regression



AFM Smith

Objective

To illustrate the Bayesian approach to fitting normal and generalized linear models.

Recommended reading

- Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society B*, **34**, 1–41.

- Broemeling, L.D. (1985). *Bayesian Analysis of Linear Models*, Marcel-Dekker.

- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapter 8.

- Wiper, M.P., Pettit, L.I. and Young, K.D.S. (2000). Bayesian inference for a Lanchester type combat model. *Naval Research Logistics*, **42**, 609–633.

Introduction: the multivariate normal distribution

Definition 22
A random variable $\mathbf{X} = (X_1, \ldots, X_k)^T$ is said to have a *multivariate normal* distribution with mean $\boldsymbol{\mu}$ and variance / covariance matrix $\boldsymbol{\Sigma}$ if

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^k.$$

In this case, we write $\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The following properties of the multivariate normal distribution are well known.

i. Any subset of $\mathbf{X}$ has a (multivariate) normal distribution.

ii. Any linear combination $\sum_{i=1}^{k} \alpha_i X_i$ is normally distributed

iii. If $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ is a linear transformation of $\mathbf{X}$, then $\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Sigma\mathbf{B}^T\right)$.

iv. If $\mathbf{X} = \begin{matrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{matrix} \Bigg| \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\begin{matrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{matrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$ then the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is

$$\mathbf{X}_1|\mathbf{x}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

The multivariate normal likelihood function

Suppose that we observe a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of data from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the likelihood function is given by

$$
\begin{aligned}
l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}) &= \frac{1}{(2\pi)^{nk/2}|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right) \\
&\propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right) \\
&\propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\left[\mathrm{tr}\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right)
\end{aligned}
$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ and $\mathrm{tr}(\mathbf{M})$ represents the trace of the matrix $\mathbf{M}$.

It is possible to carry out Bayesian inference with conjugate priors for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. We shall consider two cases which reflect different levels of knowledge about the variance-covariance matrix $\boldsymbol{\Sigma}$.

Conjugate Bayesian inference for the multivariate normal distribution I: $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$

Firstly, consider the case where the variance-covariance matrix is known up to a constant, i.e. $\boldsymbol{\Sigma} = \frac{1}{\phi}\mathbf{C}$ where $\mathbf{C}$ is a known matrix. Then, we have $\mathbf{X}|\boldsymbol{\mu}, \phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\mathbf{C}\right)$ and the likelihood function is

$$l(\boldsymbol{\mu}, \phi|\mathbf{x}) \propto \phi^{\frac{nk}{2}} \exp\left(-\frac{\phi}{2}\left[\text{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\mathbf{C}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right).$$

Analogous to the univariate case, it can be seen that a multivariate normal-gamma prior distribution is conjugate.

The multivariate normal gamma distribution

We say that $\boldsymbol{\mu}, \phi$ have a *multivariate normal gamma prior* with parameters $\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}$ if

$$\begin{aligned} \boldsymbol{\mu}|\phi &\sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\phi}\mathbf{V}\right) \\ \phi &\sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right). \end{aligned}$$

In this case, we write $\boldsymbol{\mu}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$.

The marginal distribution of $\boldsymbol{\mu}$

In this case, the marginal distribution of $\boldsymbol{\mu}$ is a multivariate, non-central t distribution.

Definition 23
A (k-dimensional) random variable, $\mathbf{T} = (T_1, \ldots, T_k)$, has a *multivariate t distribution* with parameters $d, \boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T$ if

$$f(\mathbf{t}) = \frac{\Gamma\left(\frac{d+k}{2}\right)}{(\pi d)^{\frac{k}{2}}|\boldsymbol{\Sigma}_T|^{1/2}\Gamma\left(\frac{d}{2}\right)}\left(1 + \frac{1}{d}(\mathbf{t} - \boldsymbol{\mu}_T)^T\boldsymbol{\Sigma}_T^{-1}(\mathbf{t} - \boldsymbol{\mu}_T)\right)^{-\frac{d+k}{2}}.$$

In this case, we write $\mathbf{T} \sim \mathcal{T}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T, d)$.

The following theorem gives the density of $\boldsymbol{\mu}$.

Theorem 34
Let $\boldsymbol{\mu}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then the marginal density of $\boldsymbol{\mu}$ is $\boldsymbol{\mu} \sim \mathcal{T}\left(\mathbf{m}, \frac{b}{a}\mathbf{V}, a\right)$.

Proof

$$
\begin{aligned}
p(\boldsymbol{\mu}) &= \int_0^\infty p(\boldsymbol{\mu}, \phi)\, d\phi \\
&= \int_0^\infty p(\boldsymbol{\mu}|\phi) p(\phi)\, d\phi \\
&= \int_0^\infty \frac{1}{(2\pi)^{\frac{k}{2}}|\mathbf{V}|^{\frac{1}{2}}} \exp\left(-\frac{\phi}{2}(\boldsymbol{\mu}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\mu}-\mathbf{m})\right) \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \phi^{\frac{a}{2}-1} \exp\left(-\frac{b}{2}\phi\right)\, d\phi \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\mathbf{V}|^{\frac{1}{2}}} \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \int_0^\infty \phi^{\frac{a+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\mu}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\mu}-\mathbf{m})\right]\right)\, d\phi \\
&= \frac{1}{(2\pi)^{\frac{k}{2}}|\mathbf{V}|^{\frac{1}{2}}} \frac{\left(\frac{b}{2}\right)^{\frac{a}{2}}}{\Gamma\left(\frac{a}{2}\right)} \Gamma\left(\frac{a+k}{2}\right) \left(\frac{b + (\boldsymbol{\mu}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\mu}-\mathbf{m})}{2}\right)^{-\frac{a+k}{2}}
\end{aligned}
$$

and, reordering terms proves the result. ∎

The posterior distribution of $\boldsymbol{\mu}, \phi$

Theorem 35
Let $\mathbf{X}|\boldsymbol{\mu}, \phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\mathbf{C}\right)$ and assume the prior distributions $\boldsymbol{\mu}|\phi \sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\phi}\mathbf{V}\right)$ and $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$. Then, given sample data $\mathbf{x}$, we have

$$\begin{aligned}
\boldsymbol{\mu}|\mathbf{x}, \phi &\sim \mathcal{N}\left(\mathbf{m}^\star, \frac{1}{\phi}\mathbf{V}^\star\right) \\
\phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
\mathbf{V}^\star &= \left(\mathbf{V}^{-1} + n\mathbf{C}^{-1}\right)^{-1} \\
\mathbf{m}^\star &= \mathbf{V}^\star\left(\mathbf{V}^{-1}\mathbf{m} + n\mathbf{C}^{-1}\bar{\mathbf{x}}\right) \\
a^\star &= a + nk \\
b^\star &= b + \operatorname{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} + n\bar{\mathbf{x}}^T\mathbf{C}^{-1}\bar{\mathbf{x}} - \mathbf{m}^\star\mathbf{V}^{\star -1}\mathbf{m}^\star.
\end{aligned}$$

Proof Exercise. The proof is analogous to the univariate case. ∎

A simplification

In the case where $\mathbf{V} \propto \mathbf{C}$, we have a simplified result, similar to that for the univariate case.

Theorem 36
Let $\boldsymbol{\mu}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \alpha\mathbf{C}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then,

$$\begin{aligned}
\boldsymbol{\mu}|\phi, \mathbf{x} &\sim \mathcal{N}\left(\frac{\alpha\mathbf{m} + n\bar{\mathbf{x}}}{\alpha + n}, \frac{1}{(\alpha+n)\phi}\mathbf{C}\right) \\
\phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b + \operatorname{tr}\left(\left(\mathbf{S} + \frac{\alpha n}{\alpha+n}(\mathbf{m}-\bar{\mathbf{x}})(\mathbf{m}-\bar{\mathbf{x}})^T\right)\mathbf{C}^{-1}\right)}{2}\right).
\end{aligned}$$

Proof This follows from the previous theorem substituting $\mathbf{V} = \frac{1}{\alpha}\mathbf{C}$. ∎

Results with the reference prior

Theorem 37
Given the prior $p(\boldsymbol{\mu}, \phi) \propto \frac{1}{\phi}$, then the posterior distribution is

$$\begin{aligned}
p(\boldsymbol{\mu}, \phi|\mathbf{x}) &\propto \phi^{\frac{nk}{2}-1} \exp\left(-\frac{\phi}{2}\left[\text{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right) + n(\boldsymbol{\mu}-\bar{\mathbf{x}})^T\mathbf{C}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{x}})\right]\right) \\
\boldsymbol{\mu}|\mathbf{x}, \phi &\sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n\phi}\mathbf{C}\right) \\
\phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{(n-1)k}{2}, \frac{\text{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{2}\right) \\
\boldsymbol{\mu}|\mathbf{x} &\sim \mathcal{T}\left(\bar{\mathbf{x}}, \frac{\text{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{n(n-1)k}\mathbf{C}, (n-1)k\right).
\end{aligned}$$

Proof $\boldsymbol{\mu}, \phi|\mathbf{x} \sim \mathcal{NG}\left(\bar{\mathbf{x}}, \frac{1}{n}\mathbf{C}, \frac{(n-1)k}{2}, \frac{\text{tr}\left(\mathbf{S}\mathbf{C}^{-1}\right)}{2}\right)$ and the rest follows. ∎

Conjugate inference for the multivariate normal distribution II: $\boldsymbol{\Sigma}$ unknown

In this case, it is useful to reparameterize the normal distribution in terms of the precision matrix $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1}$ when the normal likelihood function becomes

$$l(\boldsymbol{\mu}, \boldsymbol{\Phi}|\mathbf{x}) \propto |\boldsymbol{\Phi}|^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left[\mathrm{tr}\left(\mathbf{S}\boldsymbol{\Phi}\right) + n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \boldsymbol{\Phi}(\boldsymbol{\mu} - \bar{\mathbf{x}})\right]\right)$$

It is clear that a conjugate prior for $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ must take a similar form to the likelihood. This is a normal-Wishart distribution.

The normal Wishart distribution

Definition 24

A $k \times k$ dimensional symmetric, positive definite random variable $\mathbf{W}$ is said to have a *Wishart distribution* with parameters d and $\mathbf{V}$ if

$$f(\mathbf{W}) = \frac{|\mathbf{W}|^{\frac{d-k-1}{2}}}{2^{\frac{dk}{2}}|V|^{\frac{d}{2}}\pi^{\frac{k(k-1)}{4}}\prod_{i=1}^{k}\Gamma\left(\frac{d+1-i}{2}\right)}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{W}\right)\right)$$

where $d > k-1$. In this case, $E[\mathbf{W}] = d\mathbf{V}$ and we write $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$.

If $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$, then the distribution of $\mathbf{W}^{-1}$ is said to be an *inverse Wishart distribution*, $\mathbf{W}^{-1} \sim \mathcal{IW}\left(d, \mathbf{V}^{-1}\right)$ with mean $E\left[\mathbf{W}^{-1}\right] = \frac{1}{d-k-1}\mathbf{V}^{-1}$.

Theorem 38
Suppose that $\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Phi} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Phi}^{-1}\right)$ and let $\boldsymbol{\mu}|\boldsymbol{\Phi} \sim \mathcal{N}\left(\mathbf{m}, \frac{1}{\alpha}\boldsymbol{\Phi}^{-1}\right)$ and $\boldsymbol{\Phi} \sim \mathcal{W}(d, \mathbf{W})$. Then:

$$\begin{aligned}
\boldsymbol{\mu}|\boldsymbol{\Phi}, \mathbf{x} &\sim \mathcal{N}\left(\frac{\alpha\mathbf{m} + n\bar{\mathbf{x}}}{\alpha + n}, \frac{1}{\alpha + n}\boldsymbol{\Phi}^{-1}\right) \\
\boldsymbol{\Phi}|\mathbf{x} &\sim \mathcal{W}\left(d + nk, \mathbf{W}^{-1} + \mathbf{S} + \frac{\alpha n}{\alpha + n}(\mathbf{m} - \bar{\mathbf{x}})(\mathbf{m} - \bar{\mathbf{x}})^T\right)
\end{aligned}$$

Proof Exercise. ∎

We can also derive a limiting prior distribution by letting $d \to 0$ when $p(\boldsymbol{\Phi}) \propto |\boldsymbol{\Phi}|^{\frac{k+1}{2}}$ when the posterior distribution is

$$\boldsymbol{\mu}|\boldsymbol{\Phi}, \mathbf{x} \sim \mathcal{N}\left(\bar{\mathbf{x}}, \frac{1}{n}\boldsymbol{\Phi}^{-1}\right) \quad \boldsymbol{\Phi}|\mathbf{x} \sim \mathcal{W}\left(n(k-1), \mathbf{S}\right).$$

Semi-conjugate inference via Gibbs sampling

The conjugacy assumption that the prior precision of $\boldsymbol{\mu}$ is proportional to the model precision $\boldsymbol{\Sigma}$ is very strong in many cases. Often, we may simply wish to use a prior distribution of form $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$ where $\mathbf{m}$ and $\mathbf{V}$ are known and a Wishart prior for $\boldsymbol{\Phi}$, say $\boldsymbol{\phi} \sim \mathcal{W}(d, \mathbf{W})$ as earlier.

In this case, the conditional posterior distributions are

$$\begin{aligned}
\boldsymbol{\mu}|\boldsymbol{\Phi}, \mathbf{x} &\sim \mathcal{N}\left(\left(\mathbf{V}^{-1} + n\boldsymbol{\Phi}\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m} + n\boldsymbol{\Phi}\bar{\mathbf{x}}\right), \left(\mathbf{V}^{-1} + n\boldsymbol{\Phi}\right)^{-1}\right) \\
\boldsymbol{\Phi}|\boldsymbol{\mu}, \mathbf{x} &\sim \mathcal{W}\left(d + n, \mathbf{W}^{-1} + \mathbf{S} + n(\boldsymbol{\mu} - \bar{x})(\boldsymbol{\mu} - \bar{\mathbf{x}})^T\right)
\end{aligned}$$

and therefore, it is straightforward to set up a Gibbs sampling algorithm to sample the joint posterior, as in the univariate case.

Aside: sampling the multivariate normal, multivariate t and Wishart distributions

Samplers for the multivariate normal distribution (usually based on the *Cholesky decomposition*) are available in most statistical packages such as `R` or `Matlab`. Sampling the multivariate t distribution is only slightly more complicated. Assume that we wish to sample from $\mathbf{T} \sim \mathcal{T}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})$. Then from Theorem 34, the distribution of $\mathbf{T}$ is the same as the marginal distribution of $\mathbf{T}$ in the two stage model

$$\mathbf{T}|\phi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{\phi}\boldsymbol{\Sigma}\right) \quad \phi \sim \mathcal{G}\left(\frac{d}{2}, \frac{d}{2}\right).$$

Thus, sampling can be undertaken by first generating values of ϕ and then generating values of $\mathbf{T}$ from the associated normal distribution.

Sampling from a Wishart distribution can be done in a straightforward way if the degrees of freedom is a natural number. Thus, assume $\mathbf{W} \sim \mathcal{W}(d, \mathbf{V})$ where $d \in \mathbb{N}$. Then the following algorithm generates a Wishart variate.

1. Simulate $\mathbf{z}_1, \ldots, \mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$

2. Define $\mathbf{W} = \sum_{i=1}^{k} \mathbf{z}_i \mathbf{z}_i^T$.

Normal linear models

Definition 25

A *normal linear model* is of form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, and we will assume initially that $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi}\mathbf{I}\right)$.

This framework includes regression models, defining, e.g. $\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$ and $\boldsymbol{\theta} = (\alpha, \beta)^T$.

It is easy to see that for such a model, a conjugate, multivariate normal-gamma prior distribution is available.

Theorem 39

Let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and assume a multivariate normal-gamma prior distribution $\boldsymbol{\theta}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then, the predictive distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{X}\mathbf{m}, \frac{b}{a}(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{I}), a\right)$$

and the posterior distribution of $\boldsymbol{\theta}, \phi$ given $\mathbf{y}$ is

$$\begin{aligned}
\boldsymbol{\theta}, \phi|\mathbf{y} &\sim \mathcal{NG}\left(\mathbf{m}^\star, \mathbf{V}^{\star -1}, \frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where} \\
\mathbf{m}^\star &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\
\mathbf{V}^\star &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1} \\
a^\star &= a + n \\
b^\star &= b + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^\star
\end{aligned}$$

Proof First we shall prove the predictive distribution formula. We have $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and therefore, the distribution of $\mathbf{y}|\phi$ is

$$\mathbf{y}|\phi \sim \mathcal{N}\left(\mathbf{Xm}, \frac{1}{\phi}\left(\mathbf{XVX}^T + \mathbf{I}\right)\right)$$

and the joint distribution of $\mathbf{y}$ and ϕ is multivariate normal-gamma

$$\mathbf{y}, \phi \sim \mathcal{NG}\left(\mathbf{Xm}, \left(\mathbf{XVX}^T + \mathbf{I}\right)^{-1}, \frac{a}{2}, \frac{b}{2}\right)$$

and therefore, the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{Xm}, \frac{b}{a}(\mathbf{XVX}^T + \mathbf{I}), a\right).$$

Now we shall evaluate the posterior distribution

$$
\begin{aligned}
p(\boldsymbol{\theta}, \phi|\mathbf{y}) &\propto \phi^{\frac{a+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\theta}-\mathbf{m})^T\mathbf{V}^{-1}(\boldsymbol{\theta}-\mathbf{m})\right]\right) \phi^{\frac{n}{2}} \exp\left(-\frac{\phi}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right) \\
&\propto \phi^{\frac{a+n+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{X}+\mathbf{V}^{-1}\right)\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\left(\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{V}^{\star-1}\mathbf{V}^\star\left(\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + \boldsymbol{\theta}^T\mathbf{V}^{\star-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{V}^{\star-1}\mathbf{m}^\star + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m}\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b + (\boldsymbol{\theta}-\mathbf{m}^\star)^T\mathbf{V}^{\star-1}(\boldsymbol{\theta}-\mathbf{m}^\star) + \mathbf{y}^T\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^\star\mathbf{V}^{\star-1}\mathbf{m}^\star\right]\right) \\
&\propto \phi^{\frac{a^\star+k}{2}-1} \exp\left(-\frac{\phi}{2}\left[b^\star + (\boldsymbol{\theta}-\mathbf{m}^\star)^T\mathbf{V}^{\star-1}(\boldsymbol{\theta}-\mathbf{m}^\star)\right]\right)
\end{aligned}
$$

which is the kernel of the required normal-gamma distribution. ∎

Interpretation of the posterior mean

We have

$$\begin{aligned} E[\boldsymbol{\theta}|\mathbf{y}] &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\ &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right) \\ &= \left(\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{V}^{-1}\mathbf{m}\right) \end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the maximum likelihood estimator. Thus, this expression may be interpreted as a weighted average of the prior estimator and the MLE, with weights proportional to precisions, as we can recall that, conditional on ϕ, the prior variance was $\frac{1}{\phi}\mathbf{V}$ and that the distribution of the MLE from the classical viewpoint is $\hat{\boldsymbol{\theta}}|\phi \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right)$.

Relation to ridge regression

The classical least squares regression solution $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ does not exist if $\mathbf{X}^T\mathbf{X}$ is not of full rank. In this case, an often employed technique is to use *ridge regression*, see Hoerl and Kennard (1970).

The ridge regression estimator is the value of which minimizes

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^2 + \alpha^2||\boldsymbol{\theta}||^2$$

and the solution can be shown to be

$$\hat{\theta}_\alpha = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

If we use a Bayesian approach with prior, $\boldsymbol{\mu}|\phi \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\alpha\phi}\mathbf{I}\right)$, then the posterior mean is

$$E[\boldsymbol{\mu}|\mathbf{y}] = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

which is equal to the ridge regression estimate.

Example: an ANOVA type model

Example 65

Consider the ANOVA model: $y_{ij} = \theta_i + \epsilon_{ij}$ where $\epsilon_{ij}|\phi \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$, for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

Thus $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{kn_k})^T$,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_1} & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$$

and $n = \sum_{i=1}^{k} n_i$ is the model dimension.

If we use conditionally independent normal priors, $\theta_i|\phi \sim \mathcal{N}\left(m_i, \frac{1}{\alpha_i \phi}\right)$ for $i = 1, \ldots, k$ and a gamma prior $\phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$, then $\mathbf{m} = (m_1, \ldots, m_k)^T$ and $\mathbf{V} = \begin{pmatrix} \frac{1}{\alpha_1} & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{\alpha_2} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{\alpha_k} \end{pmatrix}$. Also,

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n_1 & 0 & 0 & \ldots & 0 \\ 0 & n_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & n_k \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X} + \mathbf{V}^{-1} = \begin{pmatrix} n_1 + \alpha_1 & 0 & 0 & \ldots & 0 \\ 0 & n_2 + \alpha_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & n_k + \alpha_k \end{pmatrix}$$

$$
\begin{aligned}
\left(\mathbf{X}^T\mathbf{X}+\mathbf{V}^{-1}\right)^{-1} &= \begin{pmatrix} \frac{1}{n_1+\alpha_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2+\alpha_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k+\alpha_k} \end{pmatrix} \\
\mathbf{X}^T\mathbf{y} &= (n_1\bar{y}_{1\cdot}, \dots, n_k\bar{y}_{k\cdot})^T \\
\mathbf{V}^{-1}\mathbf{m} &= (\alpha_1 m_1, \dots, \alpha_k m_k)^T \\
\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m} &= (n_1\bar{y}_{1\cdot}+\alpha_1 m_1, \dots, n_k\bar{y}_{k\cdot}+\alpha_k m_k)^T \\
\left(\mathbf{X}^T\mathbf{X}+\mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}+\mathbf{V}^{-1}\mathbf{m}\right) &= \left(\frac{n_1\bar{y}_{1\cdot}+\alpha_1 m_1}{n_1+\alpha_1}, \dots, \frac{n_k\bar{y}_{k\cdot}+\alpha_k m_k}{n_k+\alpha_k}\right)^T \quad \text{so}
\end{aligned}
$$

$$
\boldsymbol{\theta}|\mathbf{y}, \phi \sim \mathcal{N}\left(\begin{pmatrix} \frac{n_1\bar{y}_{1\cdot}+\alpha_1 m_1}{n_1+\alpha_1} \\ \frac{n_2\bar{y}_{2\cdot}+\alpha_2 m_2}{n_2+\alpha_2} \\ \vdots \\ \vdots \\ \frac{n_k\bar{y}_{k\cdot}+\alpha_k m_k}{n_k+\alpha_k} \end{pmatrix}, \frac{1}{\phi} \begin{pmatrix} \frac{1}{n_1+\alpha_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2+\alpha_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k+\alpha_k} \end{pmatrix} \right).
$$

Now we can calculate the posterior distribution of ϕ.

$$\begin{aligned}
\mathbf{y}^T\mathbf{y} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 \\
\mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} &= \sum_{i=1}^{k} \alpha_i m_i^2 \\
\mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^{\star} &= \sum_{i=1}^{k} \frac{(n_i\bar{y}_{i\cdot} + \alpha_i m_i)^2}{n_i + \alpha_i} \\
\mathbf{y}^t\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^{\star} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k}\frac{n_i\alpha_i}{n_i+\alpha_i}(\bar{y}_{i\cdot} - m_i)^2 \text{ so}
\end{aligned}$$

$$\phi|\mathbf{y} \sim \mathcal{G}\left(\frac{a+n}{2}, \frac{b + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k}\frac{n_i\alpha_i}{n_i+\alpha_i}(\bar{y}_{i\cdot} - m_i)^2}{2}\right)$$

Often we are interested in the differences in group means, e.g. $\theta_1 - \theta_2$. Here we have

$$\theta_1 - \theta_2 | \mathbf{y}, \phi \sim \mathcal{N}\left(\frac{n_1 \bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} - \frac{n_2 \bar{y}_{2\cdot} + \alpha_2 m_2}{n_2 + \alpha_2}, \frac{1}{\phi}\left(\frac{1}{\alpha_1 + n_1} + \frac{1}{\alpha_2 + n_2} \right) \right)$$

and therefore, a posterior, 95% interval for $\theta_1 - \theta_2$ is given by

$$\frac{n_1 \bar{y}_{1\cdot} + \alpha_1 m_1}{n_1 + \alpha_1} - \frac{n_2 \bar{y}_{2\cdot} + \alpha_2 m_2}{n_2 + \alpha_2} \pm \sqrt{\left(\frac{1}{\alpha_1 + n_1} + \frac{1}{\alpha_2 + n_2} \right) \frac{b + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} \frac{n_i \alpha_i}{n_i + \alpha_i} (\bar{y}_{i\cdot} - m_i)^2}{a + n}} t_{a+n}(0.975).$$

Limiting results for the linear model

Assume that we use the limiting prior $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$. Then, we have

$$
\begin{aligned}
p(\boldsymbol{\theta}, \phi|\mathbf{y}) &\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right]\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right]\right) \\
&\propto \phi^{\frac{n}{2}-1} \exp\left(-\frac{\phi}{2}\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}\right]\right) \\
\boldsymbol{\theta}|\mathbf{y}, \phi &\sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \frac{1}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}\right) \quad \text{and} \quad \phi|\mathbf{y} \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}}{2}\right) \\
\boldsymbol{\theta}|\mathbf{y} &\sim \mathcal{T}\left(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}, n-k\right) \quad \text{where } \hat{\sigma}^2 = \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\theta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}}{n-k}.
\end{aligned}
$$

Note that $\hat{\sigma}^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}})$ is the usual classical estimator of σ^2.

In this case, Bayesian credible intervals, estimators etc. will coincide with their classical counterparts.

One should note however that the propriety of the posterior distribution in this case relies on two conditions:

1. $n > k$,

2. $\mathbf{X}^T\mathbf{X}$ is off full rank.

If either of these two conditions is not satisfied, then the posterior distribution will be improper.

Example 66

In Example 65, suppose that we use the reference prior $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$. Then, we have

$$\begin{aligned}
\boldsymbol{\theta}|\mathbf{y} &\sim \mathcal{N}\left(\begin{pmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \vdots \\ \bar{y}_{k\cdot} \end{pmatrix}, \frac{1}{\phi}\begin{pmatrix} \frac{1}{n_1} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{n_2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{n_k} \end{pmatrix}\right) \\
\phi|\mathbf{y} &\sim \mathcal{G}\left(\frac{n-k}{2}, \frac{(n-k)\hat{\sigma}^2}{2}\right)
\end{aligned}$$

where $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$ is the classical variance estimate for this problem.

A 95% posterior interval for $\theta_1 - \theta_2$ is given by $\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{n-k}(0.975)$ which is equal to the usual, classical interval.

Example 67

Consider the simple linear regression model, $y_i = \alpha + \beta x_i + \epsilon_i$, for $i = 1, \ldots, n$, where $\epsilon_i \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$ and suppose that we use the limiting prior $p(\alpha, \beta, \phi) \propto \frac{1}{\phi}$. Then, we have

$$\begin{aligned}
\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \Big| \mathbf{y}, \phi &\sim \mathcal{N}\left(\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}, \frac{1}{\phi n S_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \right) \\
\phi | \mathbf{y} &\sim \mathcal{G}\left(\frac{n-2}{2}, \frac{S_{yy}(1-r^2)}{2} \right) \\
\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \Big| \mathbf{y} &\sim \mathcal{T}\left(\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}, \frac{\hat{\sigma}^2}{n} \frac{1}{S_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}, n-2 \right)
\end{aligned}$$

where $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$, $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ and $\hat{\sigma}^2 = \frac{S_{yy}(1-r^2)}{n-2}$.

Thus, the marginal distributions of α and β are

$$\begin{aligned}\alpha|\mathbf{y} &\sim \mathcal{T}\left(\hat{\alpha}, \hat{\sigma}^2\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}, n-2\right)\\ \beta|\mathbf{y} &\sim \mathcal{T}\left(\hat{\beta}, \frac{\hat{\sigma}^2}{S_{xx}}, n-2\right)\end{aligned}$$

and therefore, for example, a 95% credible interval for β is given by

$$\hat{\beta} \pm \frac{\hat{\sigma}}{\sqrt{S_{xx}}} t_{n-2}(0.975)$$

equal to the usual classical interval.

Suppose now that we wish to predict a future observation $y_{new} = \alpha + \beta x_{new} + \epsilon_{new}$. Then

$$\begin{aligned}
E[y_{new}|\phi, \mathbf{y}] &= \hat{\alpha} + \hat{\beta} x_{new} \\
V[y_{new}|\phi, \mathbf{y}] &= \frac{1}{\phi}\left(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} + \frac{x_{new}^2}{nS_{xx}} - 2\frac{x_{new}\bar{x}}{nS_{xx}} + 1\right) \\
&= \frac{1}{\phi}\left(\frac{(x_{new} - \bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right) \\
y_{new}|\mathbf{y} &\sim \mathcal{T}\left(\hat{\alpha} + \hat{\beta} x_{new}, \hat{\sigma}^2\left(\frac{(x_{new} - \bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right), n-2\right)
\end{aligned}$$

and thus, a 95% credible interval for y_{new} is

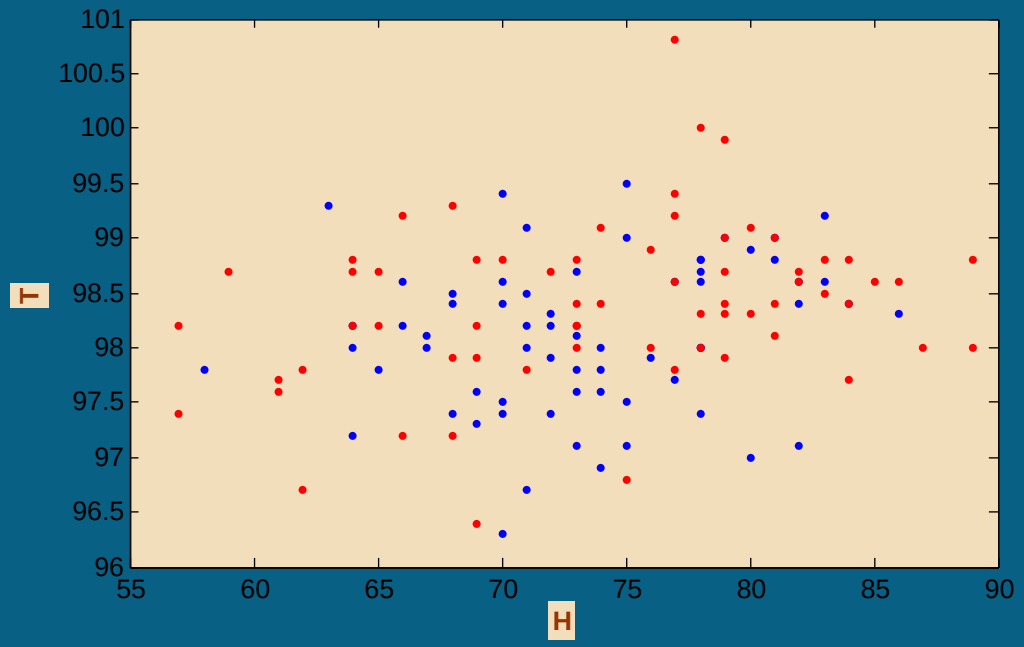
$$\hat{\alpha} + \hat{\beta} x_{new} \pm \hat{\sigma}\left(\frac{(x_{new} - \bar{x})^2}{nS_{xx}} + \frac{1}{n} + 1\right) t_{n-2}(0.975)$$

which coincides with the usual, classical interval.

The normal body temperature example again

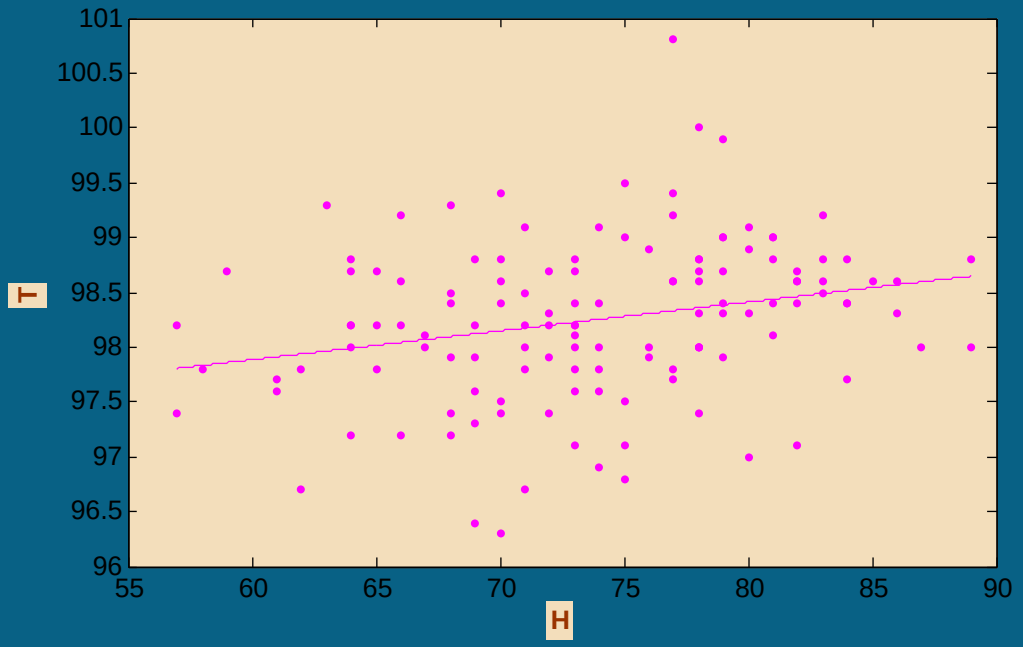
Example 68

In Example 20, we studied the average body temperatures of humans and we saw in Example 21 that these were different for men and women. Now we shall consider the effects of introducing a covariate. In the following diagram, temperature is plotted against heart rate (beats per minute) for male (blue) and female (red) subjects



It is reasonable to assume that body temperature would be related to heart rate. Thus, we might assume the global linear model, $T_i = \alpha + \beta H \epsilon_i$, independent of temperature. Fitting this model with a non-informative prior leads to the estimated regression equation

$$E[T|H, \text{data}] = 96.31 + 0.021H.$$



However, earlier we supposed that gender also influenced body temperature and therefore, a model taking gender into account might be considered. Thus, we assume

$$T_{ij} = \alpha_i + \beta H_{ij} + \epsilon_{ij}$$

where T_{ij} represents the temperature of subject j in group $i = 1$ (men) or $i = 2$ (women) and H_{ij} is the subject's heartrate and $\epsilon_{ij} \sim \mathcal{N}\left(0, \frac{1}{\phi}\right)$ is a random error.

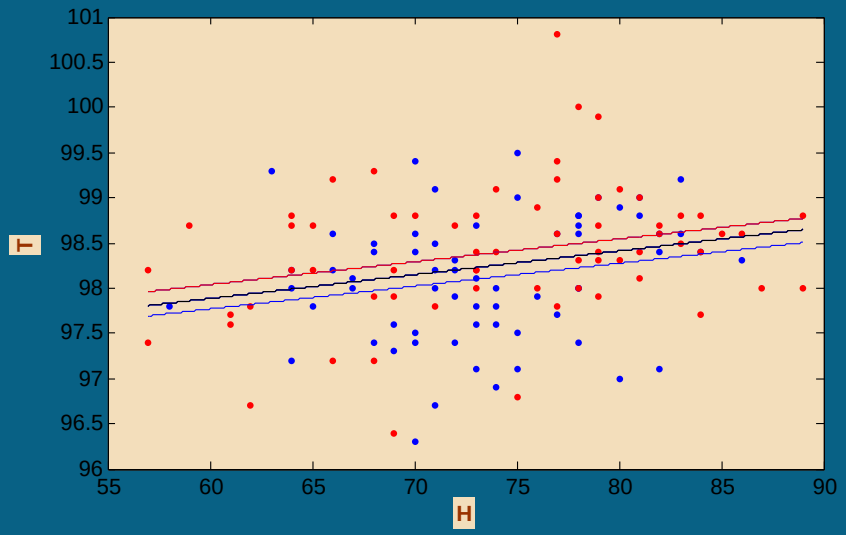
Representing this model in linear form, we have $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta)^T$, $\mathbf{y} = (T_{11}, \ldots, T_{1n_1}, T_{21}, \ldots, T_{2n_2})^T$ where n_1 and n_2 are the numbers of men and women sampled respectively and

$$\mathbf{X}^T = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \\ H_{11} & \ldots & H_{1n_1} & H_{21} & \ldots & H_{2n_2} \end{pmatrix}$$

Assume that we use the prior $p(\alpha_1, \alpha_2, \beta, \phi)$ posterior mean parameter values for $\boldsymbol{\alpha}, \beta$ are

$$\begin{aligned} E[\alpha_1|\mathbf{T}] &= 96.2508 \\ E[\alpha_2|\mathbf{T}] &= 96.5202 \\ E[\beta|\mathbf{T}] &= 0.0253 \end{aligned}$$

The posterior distribution of ϕ is $\phi|\mathbf{y} \sim \mathcal{G}\left(\frac{127}{2}, \frac{62.5}{2}\right)$. The following diagram shows the results of fitting the simple and combined models.



It is interesting to assess whether or not the difference between the sexes is still important. Thus, we can calculate the posterior distribution of $\alpha_1 - \alpha_2$. We have $\alpha_1 - \alpha_2 | \mathbf{y}, \phi \sim \mathcal{N}\left(-0.2694, \frac{0.031}{\phi}\right)$ and therefore a 95% posterior credible interval is

$$-0.2694 \pm \sqrt{0.031 * 62.5/127} * t_{127}(0 - 975) = (-0.5110, -0.0278).$$

Thus, it seems likely that the combined model is superior to the simple regression model.

Note that in order to undertake a formal analysis of this, we could use fractional or intrinsic Bayes factors as the prior distributions in this case were improper.

Including covariance in the linear model

It is possible to fit a more general linear model where we do not assume that the model variance is proportional to the identity. One possibility is to assume that the model variance is proportional to a known matrix ($\mathbf{C}$). Lindley and Smith (1972) then demonstrate the following theorem.

Theorem 40
Let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}|\phi \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi}\mathbf{C}\right)$ with prior distribution $\boldsymbol{\theta}, \phi \sim \mathcal{NG}\left(\mathbf{m}, \mathbf{V}^{-1}, \frac{a}{2}, \frac{b}{2}\right)$. Then, the predictive distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{T}\left(\mathbf{X}\mathbf{m}, \frac{b}{a}\left(\mathbf{X}\mathbf{V}\mathbf{X}^T + \mathbf{C}\right), a\right)$$

and the posterior distribution of $\boldsymbol{\mu}, \phi|\mathbf{y}$ is

$$\boldsymbol{\theta}|\mathbf{y}, \phi \sim \mathcal{N}\left(\mathbf{m}^\star, \frac{1}{\phi}\mathbf{V}^\star\right) \quad \phi|\mathbf{y} \sim \mathcal{G}\left(\frac{a^\star}{2}, \frac{b^\star}{2}\right) \quad \text{where}$$

$$
\begin{aligned}
\mathbf{m}^\star &= \left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X} + \mathbf{V}^{-1}\right)^{-1}\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}\right), \\
\mathbf{V}^\star &= \left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X} + \mathbf{V}^{-1}\right)^{-1} \\
a^\star &= a + n \\
b^\star &= b + \mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} + \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}^{\star T}\mathbf{V}^{\star -1}\mathbf{m}^\star
\end{aligned}
$$

Proof Exercise. ∎

In the limiting case, given the prior $p(\boldsymbol{\theta}, \phi) \propto \frac{1}{\phi}$, it is easy to show that these posterior distributions converge to produce posterior distributions which lead to the same numerical results as in standard classical inference, e.g.

$$
\boldsymbol{\theta}|\mathbf{y}, \phi \sim \mathcal{N}\left(\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{C}^{-1}\mathbf{y}, \frac{1}{\phi}\left(\mathbf{X}^T\mathbf{C}^{-1}\mathbf{X}\right)^{-1}\right)
$$

and the posterior mean is the classical MLE of $\boldsymbol{\theta}$.

The SUR model

A more general approach is to assume an unknown model variance-covariance matrix $\mathbf{\Sigma}$ and set an inverse Wishart prior distribution, e.g. $\mathbf{\Phi} = \mathbf{\Sigma}^{-1} \sim \mathcal{W}(d, \mathbf{W})$.

Example 69

Seemingly unrelated regression (SUR) is a well known econometric model. In the traditional SUR model, we have M equations of form

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, \ldots, M$ where $\mathbf{y}_i$ is an N dimensional vector of observations on a dependent variable, $\mathbf{X}_i$ is a $(N \times K_i)$ matrix of observations on K_i independent variables, $\boldsymbol{\beta}_i$ is a K_i dimensional vector of unknown regression coefficients and $\boldsymbol{\epsilon}_i$ is an N dimensional, unobserved error vector.

The M equations can be written as

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & & & \\ & \mathbf{X}_2 & & \\ & & \ddots & \\ & & & \mathbf{X}_M \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{pmatrix}$$

and written compactly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ has dimension $(NM \times 1)$, $\mathbf{X}$ has dimension $(NM \times K)$ where $K = \sum_{i=1}^{M} K_i$, $\boldsymbol{\beta}$ is $(K \times 1)$ and $\boldsymbol{\epsilon}$ has dimension $(NM \times 1)$. Assume the distribution of $\boldsymbol{\epsilon}$ is

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_N\right).$$

The likelihood function is now

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\mathbf{y}) &\propto \boldsymbol{\Sigma}^{-\frac{N}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \\ &\propto \boldsymbol{\Sigma}^{-\frac{N}{2}} \exp\left(-\frac{1}{2} tr(\mathbf{A}\boldsymbol{\Sigma}^{-1})\right) \end{aligned}$$

where $\mathbf{A}$ is a $(M \times M)$ matrix with (i, j)'th element

$$(\mathbf{A})_{ij} = (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i)^T (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}_j)$$

A natural, uninformative prior is $p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{M+1}{2}}$ and then, given this prior, we have immediately that

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\Sigma} &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \left[\mathbf{X}^T(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{X}\right]^{-1}\right) \\ \boldsymbol{\Sigma} &\sim \mathcal{IW}(N, \mathbf{A}) \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{X}\right]^{-1} \mathbf{X}^T(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N)\mathbf{y}$ is the standard, least squares estimator.

The three stage linear model and ideas of hierarchical models

Up to now, we have used direct priors on the regression parameters $\boldsymbol{\theta}$, ϕ. In some cases, it may be more appropriate to use *hierarchical priors*. One example is the three stage linear model of Lindley and Smith (1972) who propose the structure

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ \boldsymbol{\theta} &\sim \mathcal{N}(\mathbf{A}\boldsymbol{\beta}, \mathbf{V}) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{m}, \mathbf{W}) \end{aligned}$$

so that the prior distribution of $\boldsymbol{\theta}$ is defined hierarchically in two stages. Lindley and Smith (1972) demonstrate how to calculate the posterior distribution in this case when all the variances are known.

Example 70

In Example 65, we assumed direct, independent priors on the group mean parameters θ_i. Often however, we may have little information about these paramaters except that we have no prior evidence that they are different. This would suggest the use of a hierarchical prior, for example,

$$\begin{aligned} \theta_i|\theta_0, \phi &\sim \mathcal{N}\left(\theta_0, \frac{1}{\alpha\phi}\right) \\ p(\theta_0, \phi) &\propto \frac{1}{\phi}. \end{aligned}$$

We shall illustrate how Bayesian inference using hierarchical priors can be carried out in the following chapter.

Generalized linear models

The generalized linear model (Nelder and Wedderburn 1972) generalizes the normal linear model by allowing the possibility of non-normal error distributions and by allowing for a non-linear relationship between $\mathbf{y}$ and $\mathbf{x}$.

Definition 26
A generalized linear model is specified by two functions:

i a conditional, exponential family density function of y given $\mathbf{x}$, parameterized by a mean parameter, $\mu = \mu(\mathbf{x}) = E[Y|\mathbf{x}]$ and (possibly) a dispersion parameter, $\phi > 0$ that is independent of $\mathbf{x}$,

ii a (one-to-one) *link function*, g, which relates the mean, $\mu = \mu(\mathbf{x})$ to the covariate vector, $\mathbf{x}$, as $g(\mu) = \mathbf{x}\boldsymbol{\theta}$.

Example 71
The logistic regression model is often used for predicting the occurrence of an event given covariates. It is assumed that

$$\begin{aligned} Y_i|p_i &\sim \mathcal{BI}(n_i, p_i) \quad \text{for } i = 1, \ldots, m, \text{ and} \\ \log \frac{p_i}{1-p_i} &= \mathbf{x}_i\boldsymbol{\theta} \end{aligned}$$

Example 72
In Poisson regression, it is supposed that

$$\begin{aligned} Y_i|\lambda_i &\sim \mathcal{P}(\lambda_i) \\ \log \lambda_i &= \mathbf{x}_i\boldsymbol{\theta} \end{aligned}$$

In both examples, we have assumed the *canonical link function* which is the natural parameterization to leave the exponential family distribution in canonical form.

The Bayesian specification of a GLM is completed by defining (typically normal or normal gamma) prior distributions $p(\boldsymbol{\theta}, \phi)$ over the unknown model parameters. As with standard linear models, when improper priors are used, it is then important to check that these lead to valid posterior distributions.

Clearly, these models will not have conjugate posterior distributions, but, usually, they are easily handled by Gibbs sampling.

In particular, the posterior distributions from these models are usually log concave and are thus easily sampled via adaptive rejection sampling, see e.g. Gilks and Wild (1992).

Example 73

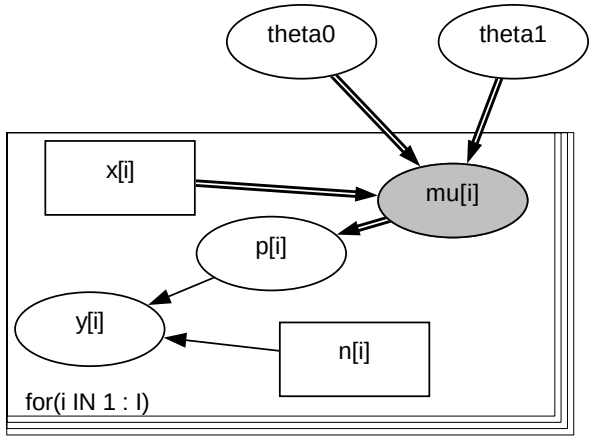
The table shows the relationship, for 64 infants, between gestational age of the infant (in weeks) at the time of birth (x) and whether the infant was breast feeding at the time of release from hospital (y).

x	28	29	30	31	32	33
$\#y = 0$	4	3	2	2	4	1
$\#y = 1$	2	2	7	7	16	14

Let x_i represent the gestational age and n_i the number of infants with this age. Then we can model the probability that y_i infants were breast feeding at time of release from hospital via a standard binomial regression model.

It is easy to set this model up in Winbugs.

name: mu[i] type: logical link: identity
value: theta0+theta1*x[i]



The doodle gives the following model:

```
model; {
   theta0 ~ dnorm( 0.0,1.0E-6)
   theta1 ~ dnorm( 0.0,1.0E-6)
   for( i in 1 : I ) {
      mu[i] <- theta0+theta1 * x[i]
   }
   for( i in 1 : I ) {
      logit(p[i]) <- mu[i]
   }
   for( i in 1 : I ) {
      y[i] ~ dbin(p[i],n[i])
   }
}
```

and we can read in the data using

```
list(I=6,x=c(28,29,30,31,32,33),n=c(6,5,9,9,20,15),
y=c(2,2,7,7,16,14))
```

In this case, we set the initial values to `theta0=theta1=1`. Given 30000 iterations to burn in and 30000 in equilibrium we have the following posterior estimates for `p` and `theta`.

node	mean	sd	MC error	2.5%	median	97.5%
p[1]	0.3783	0.1376	0.008533	0.1332	0.3714	0.6544
p[2]	0.5089	0.1118	0.00595	0.2842	0.5117	0.7173
p[3]	0.646	0.07742	0.002298	0.484	0.6501	0.7858
p[4]	0.7636	0.0577	9.063E-4	0.6435	0.7667	0.8673
p[5]	0.8483	0.05263	0.002359	0.7332	0.8528	0.9374
p[6]	0.9032	0.04844	0.002742	0.789	0.9108	0.9747
theta0	-16.85	6.479	0.4591	-30.71	-16.42	-5.739
theta1	0.5823	0.2112	0.01497	0.2222	0.567	1.036

The posterior mean values are quite close to the sample proportions.

Application V: Inference for Lanchester's combat models

Lanchester

Lanchester (1916) developed a system of equations for modeling the losses of combating forces.

Lanchester's equations

The Lanchester equations for modern warfare (aimed combat, without reinforcements) between armies of size $x(t)$ and $y(t)$ are

$$\begin{aligned} \frac{\partial x}{\partial t} &= -\alpha y \\ \frac{\partial y}{\partial t} &= -\beta x \end{aligned}$$

for $x(t), y(t) > 0$.

These equations lead to the well-known Lanchester square law

$$\alpha \left(x(0)^2 - x(t)^2 \right) = \beta \left(y(0)^2 - y(t)^2 \right).$$

This law has been fitted to some real combat situations such as the battle of Iwo Jima (Engel 1954).

Different types of warfare

A more general system of differential equations which includes the square law can be used to represent different forms of warfare as follows

$$\begin{aligned} \frac{\partial x}{\partial t} &= -\beta x^{\phi_1} y^{\phi_2} \\ \frac{\partial y}{\partial t} &= -\alpha y^{\phi_1} x^{\phi_2} \end{aligned}$$

Here, $\boldsymbol{\phi} = (0, 1)$ gives the square law, $\boldsymbol{\phi} = (1, 1)$ gives a linear law representing unaimed fire combats, $\boldsymbol{\phi} = (0, 0)$ leads to a different linear law representing hand-to hand combats and $\boldsymbol{\phi} = (1, 0)$ leads to a logistic law which has been used to represent large scale combats such as the American Civil War. See e.g. Weiss (1966).

Introducing uncertainty

In modeling combat situations, interest lies in:

- classifying battle types (historical analysis),
- assessing relative fighting strengths of the two armies (i.e. the ratio α/β),
- predicting casualties,
- predicting who will win the battle.

However, the basic Lanchester models are deterministic and thus, for instance, the battle winner is predetermined given the model parameters. Thus, we need to introduce a random element into the Lanchester systems.

One possibility is to consider stochastic Lanchester models based on Poisson process type assumptions, see e.g. Clark (1969), Goldie (1977) or Pettit et al (2003). Following Wiper et al (2000), an alternative is to discretize time, linearize the Lanchester systems and fit a regression model.

Discretization and Linearization

Given daily casualty data, we can discretize the Lanchester equations to give:

$$\Delta x_t \approx \beta x_{t-1}^{\phi_1} y_{t-1}^{\phi_2} \quad \Delta y_t \approx \alpha y_{t-1}^{\phi_1} x_{t-1}^{\phi_2}$$

where Δx_t and Δy_t are the daily casualties recorded in the two armies.

Bracken (1995) attempted to fit this model directly to combat data from the Ardennes campaign. However, it seems more natural to linearize the model. Taking logs and introducing an error term, we have:

$$\mathbf{z}_t = \boldsymbol{\theta} + \mathbf{P}_{t-1}\boldsymbol{\phi} + \boldsymbol{\epsilon}_t$$

where $\mathbf{z}_t$ are the logged casualties of the two armies on day t, $\boldsymbol{\phi} = (\phi_1, \phi_2)^T$ and $\mathbf{P}_{t-1} = \begin{pmatrix} \log y_{t-1} & \log x_{t-1} \\ \log x_{t-1} & \log y_{t-1} \end{pmatrix}$.

Assuming that the error distribution is normal, $\boldsymbol{\epsilon}_t | \boldsymbol{\Phi} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Phi}^{-1}\right)$, then we have a (multivariate) linear model.

Analysis of the Ardennes campaign data

The Battle of the Bulge

The *Battle of the Bulge* in 1944, was one of the largest battles in the Second World War involving, for example, over 250000 US troops with nearly 20000 US casualties and even more German casualties. The Germans launched an initial surprise offensive under cloud cover and attacked during 5 days when the allied troops mounted a counterattack which lead to the eventual defeat of the German army 23 days later.

For a full description, see:

http://en.wikipedia.org/wiki/Battle_of_the_Bulge

The forces involved in the battle were troops, cannons and tanks and for the purposes of this analysis, we consider a composite force for each army which is a weighted measure of these components.

As it seems likely that casualties may be affected by whether an army is attacking or not, the general regression model was modified to include a factor $\boldsymbol{\delta}$ which indicated which army was attacking so that the full model is

$$\mathbf{z}_t = \boldsymbol{\theta} + \mathbf{P}_{t-1}\boldsymbol{\phi} + \mathbf{I}_t\boldsymbol{\delta} + \boldsymbol{\epsilon}_t.$$

Here $\mathbf{I}_t = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ if the Germans were attacking on day t and $\mathbf{I}_t = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ if the Americans were attacking.

Relatively informative prior distributions were used, with a Wishart prior structure for Φ. The model was then fitted using Gibbs sampling. The posterior mean parameter estimates are given below.

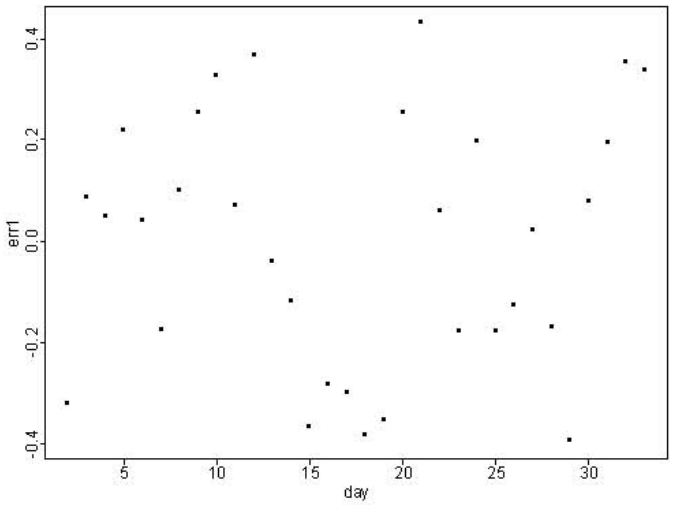
Parameter	Mean (standard deviation)	
$\boldsymbol{\theta}$	$7.88\ (.09)$	$8.04\ (.08)$
$\boldsymbol{\phi}$	$0.84\ (.41)$	$0.27\ (.40)$
$\boldsymbol{\delta}$	$-0.38\ (.09)$	$-0.01\ (.09)$

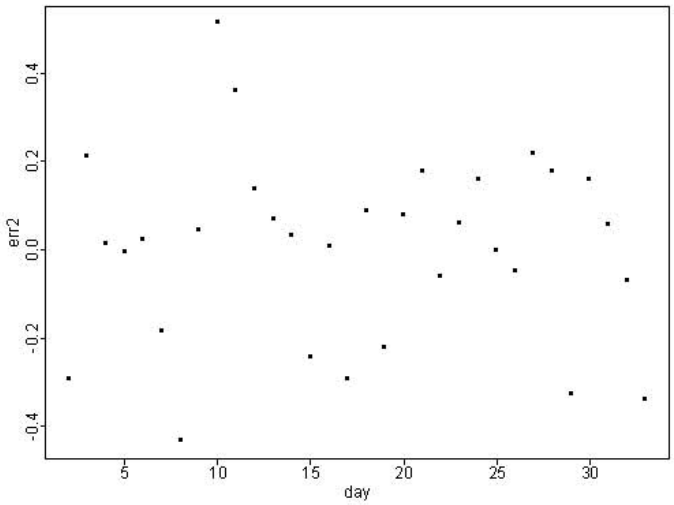
There is no clear evidence in favour of any one of the standard Lanchester laws as against any other although, using Bayes factors to compare the specific models $\boldsymbol{\phi} = (0,1)$, $(0,0)$, $(1,0)$ and $(1,1)$ suggests that the logistic law, $\boldsymbol{\phi} = (1,0)$ is the most likely.

There is strong evidence that the American casualties were typically higher when the Germans were attacking although the German casualties did not seem to be influenced by this.

Goodness of fit

In order to assess the fit of the model, we calculated the predicted errors $\mathbf{z}_t - E[\mathbf{z}_t|\text{data}]$. A plot of these errors for the first American (left) and German (right) armies is given below.





There is slight evidence of lack of fit for the German forces. More complex models might be preferable. See e.g. Lucas and Turkes (2003).

Conclusions and Extensions

- Bayesian inference for Lanchester models is straightforward to implement.
- There is high colinearity in these models and proper prior information is necessary.
- Many models fit the Ardennes data almost equally well.
- The Lanchester model has been proposed as a model for competition between ants, see e.g. McGlynn (2000) and for business competition, see e.g. Campbell and Roberts (2006).
- The Lanchester equations are similar to the Lotka-Volterra equations for predator prey systems. It is possible to extend the approach to these systems.

References

Bracken, J. (1995). Lanchester models of the Ardennes campaign. *Naval Research Logistics*, **42**, 559-577.

Campbell, N.C.G. and Roberts, K.J. (2006). Lanchester market structures: A Japanese approach to the analysis of business competition. *Strategic Management Journal*, **7**, 189–200.

Clark, G.M. (1969). *The combat analysis model*. Ph.D. thesis, The Ohio State University.

Engel, J.H. (1954). A verification of Lanchester's law. *Operations Research*, **2**, 53–71.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Goldie, C.M. (1977). Lanchester square-law battles: Transient and terminal distributions. *Journal of Applied Probability*, **14**, 604-610.

Hoerl, A.E. and Kennard R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Lanchester, F.W. (1916). *Aircraft in warfare the dawn of the fourth arm*. London: Constable.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society Series B*, **34**, 1-41.

Lucas, T.W. and Turkes, T. (2003). Fitting Lanchester equations to the battles of Kursk and Ardennes. *Naval Research Logistics*, **51**, 95–116.

McGlynn, T.P. (2000). Do Lanchester's laws of combat describe competition in ants? *Behavioral Ecology*, **11**, 686–690.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, **132**, 107–120.

Pettit, L.I., Wiper, M.P. and Young, K.D.S. (2003). Bayesian inference for some Lanchester combat laws. *European Journal of Operational Research*, **148**, 152–165.

Weiss, H.K. (1966). Combat models and historical data: The US Civil War. *Operations Research*, **14**, 759-790.

Wiper, M.P., Pettit, L.I. and Young, K.D.S. (2000). Bayesian inference for a Lanchester type combat model. *Naval Research Logistics*, **42**, 609-633.

10. Exchangeability and hierarchical models

Objective

Introduce exchangeability and its relation to Bayesian hierarchical models. Show how to fit such models using fully and empirical Bayesian methods.

Recommended reading

- Bernardo, J.M. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, **4**, 111–121. Available from

  http://www.uv.es/~bernardo/Exchangeability.pdf

- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, **39**, 83–87.

- Yung, K.H. (1999). Explaining the Stein paradox. Available from

  http://www.cs.toronto.edu/~roweis/csc2515/readings/stein_paradox.pdf

Exchangeability

Suppose that we have a sequence of variables $X_1, X_2, \ldots, X_n$. Then, in many cases, we might wish to assume that the subscripts of each individual variable are uninformative. For example, in tossing a coin three times, it is natural to assume that $P(0,0,1) = P(0,1,0) = P(1,0,0)$. This idea underlines the concept of *exchangeability* as developed by De Finetti (1970, 1974).

Definition 27
A sequence of random variables $X_1, \ldots, X_n$ is said to be (finitely) *exchangeable* if the distribution of any permutation is the same as that of any other permutation, that is if

$$P(\cap_{i=1}^{n} X_{\pi(i)}) = P(\cap_{i=1}^{n} X_i)$$

for all permutation functions $\pi(\cdot)$.

The definition of exchangeability can be extended to infinite sequences of variables.

Definition 28
An infinite sequence, $X_1, X_2, \ldots$ is said to be (infinitely) exchangeable if every finite subsequence is judged to be exchangeable in the above sense.

Thus, a sequence of variables that are judged to be independent and identically distributed is exchangeable. However, exchangeability is clearly a weaker concept more related to symmetry.

For example, if $X_1, X_2, \ldots, X_5$ are the results of 5 draws without replacement from a pack of cards, then this sequence is exchangeable but clearly, the variables are not independent.

Typical non exchangeable variables are Markov chains or other time varying sequences.

De Finetti's theorem for 0-1 random variables

Assume that we have an infinitely exchangeable sequence of 0-1 variables. For example, we may believe that an infinite sequence of tosses of the same coin is exchangeable. Then, De Finetti derived the following theorem.

Theorem 41
If $X_1, X_2, \ldots$ is any infinitely exchangeable sequence of 0-1 variables with probability measure F then their exists a distribution function P such that

$$f(x_1, \ldots, x_m) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{m-x_i} \, dP(\theta)$$

where $P(\theta) = \lim_{n\to\infty} F(Y_n/n \leq \theta)$, $Y_n = \sum_{i=1}^{n} X_i$ and $\lim_{n\to\infty} Y_n/n = \theta$.

Proof See Bernardo and Smith (1994). ∎

Interpretation of De Finetti's theorem

If a sequence $X_1, X_2, \ldots$ of 0-1 variables is judged to be exchangeable, then we may interpret this as if

- The X_i are judged to be Bernoulli variables given some random variable θ.
- θ is given a probability distribution P.
- Using the strong law of large numbers, $\theta = \lim_{n\to\infty} Y_n/n$ which implies that we can interpret P as representing our beliefs about the limiting frequency of 1's.

Thus, P may be interpreted as a prior distribution.

http://en.wikipedia.org/wiki/De_Finetti%27s_theorem

An immediate consequence of Theorem 41 is that if we define $Y_n = \sum_{i=1}^{n} X_i$, then automatically, the distribution of Y_n can be represented as

$$f(Y_n = y_n) = \int_0^1 \left(\begin{array}{c} n \\ y_n \end{array} \right) \theta^{y_n} (1-\theta)^{n-y_n} P(\theta) d\theta.$$

Thus, if we are expressing our beliefs about Y_n, then we are justified in acting as if the likelihood were binomial and with a prior distribution $P(\theta)$.

However, a much stronger general representation theorem is available for any infinitely exchangeable sequence of variables.

De Finetti's general theorem

Theorem 42
Let $X_1, X_2, \ldots$ be an infinitely exchangeable sequence of variables with probability measure F. Then, their exists a probability measure P such that the joint distribution of $X_1, \ldots, X_n$ can be represented as

$$F(x_1, \ldots, x_n) = \int_{\mathcal{G}} \prod_{i=1}^{n} G(x_i) dP(G)$$

where $\mathcal{G}$ is the space of distribution functions, $P(G) = \lim_{n\to\infty} F(G_n)$ and G_n is the empirical distribution function defined by $X_1, \ldots, X_n$.

Proof See Bernardo and Smith (1994). ∎

The theorem implies that if the X_i are judged to be exchangeable, then their exists a variable $\boldsymbol{\theta}$ such that

$$F(\mathbf{x}) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{n} F(x_i|\boldsymbol{\theta}) dP(\boldsymbol{\theta}).$$

De Finetti's theorems provide a theoretical justification of Bayesian inference based on the assumptions of exchangeability. However, they are generally not very useful in the practical determination of the form of the prior distribution.

Certain extensions can be used to justify more specific distributional models. For example if we believe that the sequence $X_1, X_2, \ldots$ is exchangeable and spherically symmetric, i.e. that the distribution of $\mathbf{X}_n$ is the same as the distribution of $\mathbf{A}\mathbf{X}_n$ for any orthogonal matrix $\mathbf{A}$, then this implies that the X's may be interpreted as normally distributed given a prior distribution on the precision.

See Bernardo and Smith (1994) for details and extensions to other models.

Hierarchical models

In many models we are unclear about the extent of our prior knowledge. Suppose we have data $\mathbf{x}$ with density $f(\mathbf{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Often we may make extra assumptions about the structural relationships between the elements of $\boldsymbol{\theta}$.

Combining such structural relationships with the assumption of exchangeability leads to the construction of prior density, $p(\boldsymbol{\theta}|\boldsymbol{\phi})$, for $\boldsymbol{\theta}$ which depends upon a further, unknown hyperparameter $\boldsymbol{\phi}$.

In such cases, following Good (1980), we say that we have a hierarchical model.

Examples of hierarchical models

Example 74

Various individuals $i = 1, \ldots, n$, take an IQ test where it is supposed that the result is

$$Y_i | \theta_i, \phi \sim \mathcal{N}\left(\theta_i, \frac{1}{\phi}\right)$$

where the outcome for subject i is supposed to depend on his or her true IQ θ_i. Now if we suppose that the true IQ's of the people in the study are exchangeable then we might reasonably assume that

$$\theta_i | \mu, \psi \sim \mathcal{N}\left(\mu, \frac{1}{\psi}\right)$$

where the unknown hyperparameters are μ, representing the mean true IQ in the population, and ψ.

Example 75

George et al (1993) analyzed data concerning failures of 10 power plant pumps. The number of failures X_i at plant i was assumed to follow a Poisson distribution

$$X_i|\theta_i \sim \mathcal{P}(\theta_i t_i) \quad \text{for } i = 1, \ldots, 10,$$

where θ_i is the failure rate for pump i and t_i is the length of operation time of the pump (in 1000s of hours).

It is natural to assume that the failure rates are exchangeable and thus we might model

$$\theta_i|\alpha, \beta \sim \mathcal{G}(\alpha, \beta)$$

where α and β are the unknown hyperparameters.

Fitting hierarchical models

The most important problem in dealing with hierarchical models is how to treat the hyperparameters ϕ. Usually, there is very little prior information available with which to estimate ϕ.

Thus, two main approaches have developed:

- The natural Bayesian approach is to use relatively uninformative prior distributions for the hyperparameters ϕ and then perform a fully Bayesian analysis.

- An alternative is to estimate the hyperparameters using classical statistical methods.

This second method is the so-called *empirical Bayes* approach which we shall explore below.

The empirical Bayesian approach

Robbins

This approach is originally due to Robbins (1955). For a full review see

http://en.wikipedia.org/wiki/Empirical_Bayes_method

Suppose that we have a model $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ with a hierarchical prior $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ where $\boldsymbol{\phi}$ is a hyperparameter. Then conditional on $\boldsymbol{\phi}$, we have from Bayes theorem that

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi}).$$

Now suppose that we do not wish to specify a hyperprior distribution for $\boldsymbol{\phi}$. Then, the *empirical Bayes* (EB) approach is to use the data to estimate $\boldsymbol{\phi}$ (via e.g. maximum likelihood, the method of moments or some alternative approach). Then the analysis proceeds as if $\boldsymbol{\phi}$ were known so that given an estimate, $\hat{\boldsymbol{\phi}}$, of the hyperparameter, then we approximate the posterior distribution of $\boldsymbol{\theta}$ by

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\hat{\boldsymbol{\phi}}).$$

Multivariate normal example

Example 76
Suppose that

$$\begin{aligned} X_i|\theta_i &\sim \mathcal{N}(\theta_i, 1) \\ \theta_i|\tau &\sim \mathcal{N}(0, \tau^2) \end{aligned}$$

for $i = 1, \ldots, n$.

Then, it is easy to see that a posteriori, $\theta_i|x_i, \tau \sim \mathcal{N}((1-B)x_i, 1-B)$ where $B = \frac{1}{1+\tau^2}$ with posterior mean $E[\theta_i|x_i, \tau] = (1-B)x_i$.

When τ is unknown, this posterior mean estimate is unsatisfactory as it depends on τ. One possibility is thus to estimate τ from the data.

In order to do this, note first that

$$\mathbf{X}|\tau \sim \mathcal{N}\left(\mathbf{0}, \left(1+\tau^2\right)\mathbf{I}_n\right)$$

and consider $S = \sum_{i=1}^{n} X_i^2$. From a classical viewpoint, we have $S\frac{1}{B}\chi_n^2$ so that if we define

$$\hat{B} = \frac{n-2}{S}$$

it is easy to show that $E[\hat{B}] = B$ so that $\hat{B}$ is an unbiased estimator of B.

Substituting $\hat{B}$ for B, we now proceed as if B were known. Therefore, we have

$$\theta_i|\mathbf{x}, (B = \hat{B}) \sim \mathcal{N}\left(\left(1-\hat{B}\right)x_i, \left(1-\hat{B}\right)\right)$$

and therefore

$$E[\theta_i|\mathbf{x}] = \left(1 - \frac{n-2}{\sum_{i=1}^{n} x_i^2}\right)x_i$$

is an EB estimate of the posterior mean.

Stein's paradox and the James Stein estimator

From a classical viewpoint, as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_n)$ it would appear that a natural estimator of $\boldsymbol{\theta}$ would be $\mathbf{X}$ itself.

However Stein (1955) showed that $\mathbf{X}$ is not even an *admissable* estimator of $\boldsymbol{\theta}$ when $n \geq 3$. James and Stein (1960) showed that an estimator which *dominates* $\mathbf{X}$ is

$$\left(1 - \frac{n-2}{\sum_{i=1}^{n} X_i^2}\right) \mathbf{X}$$

which is exactly the EB estimator we have just derived.

For a fuller discussion of Stein's paradox, see Yung (1999).

Example 77

The batting average of a baseball player is the number of hits S_i divided by the number of times at bat. Supposing that n players have each gone out to bat k times, then (assuming exchangeability of hits for each player) the batting average of the i'th player is $S_i/k \sim \mathcal{B}(k, p_i)$ where p_i is the probability that they hit the ball on a given time at bat.

Then, using an arc sin transformation,

$$\begin{aligned} X_i &= 2\sqrt{k}\sin^{-1}\sqrt{\frac{S_i}{k}} \\ \theta_i &= 2\sqrt{k}\sin^{-1}\sqrt{p_i} \end{aligned}$$

we have that approximately, $X_i|\theta_i \sim \mathcal{N}(\theta_i, 1)$. Now assuming exchangeability of players, it is reasonable to assume that $\theta_i|\mu, \sigma \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

The empirical Bayes approach is now to estimate μ and σ from the data. Noting that the marginal distribution of X_i given the hyperparameters is $X_i|\mu,\sigma \sim \mathcal{N}\left(\mu, 1+\sigma^2\right)$ then,

$$E[\bar{X}|\mu,\sigma] = \mu \quad E\left[\frac{(n-3)}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right] = \frac{1}{1+\sigma^2}$$

so we have method of moment estimates $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-3} - 1$.

Now, given $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, we have $\theta_i|\mathbf{x} \sim \mathcal{N}\left(\left(1+\frac{1}{\hat{\sigma}^2}\right)^{-1}\left(x_i + \frac{1}{\hat{\sigma}^2}\bar{x}\right), \left(1+\frac{1}{\hat{\sigma}^2}\right)^{-1}\right)$.

Thus, the EB estimator for θ_i takes the form $\hat{\theta}_i^{EB} = \left(1+\frac{1}{\hat{\sigma}^2}\right)^{-1}\left(x_i + \frac{1}{\hat{\sigma}^2}\bar{x}\right)$ and inverting of the arc sin law transformation, we have $\hat{p}_i^{EB} = \sin^2\left(\frac{\hat{\theta}_i^{EB}}{2\sqrt{n}}\right)$.

Efron and Morris (1975) analyzed the batting averages of $n = 18$ baseball players over their first 45 at bats and over the remainder of the season. A table of the data follow.

Name	$\hat{p}_i$(1st 45)	p_i(Rest)	$\hat{p}_i^{EB}$
Clemente	0.400	0.346	0.290
F.Robinson	0.378	0.298	0.286
F.Howard	0.356	0.276	0.282
Johnstone	0.333	0.222	0.277
Berry	0.311	0.273	0.273
Spencer	0.311	0.270	0.273
Kessinger	0.289	0.263	0.268
Alvarado	0.267	0.210	0.264
Santo	0.244	0.269	0.259
Swoboda	0.244	0.230	0.259
Unser	0.222	0.264	0.254
Williams	0.222	0.256	0.254
Scott	0.222	0.303	0.254
Petrocelli	0.222	0.264	0.254
Rodriguez	0.222	0.226	0.254
Campaneris	0.200	0.285	0.249
Munson	0.178	0.316	0.244
Alvis	0.156	0.200	0.239

It can be seen immediately that the EB estimates correspond more closely to the true batting averages over the season than do the raw estimates.

Criticisms and characteristics of the empirical Bayes approach

In general, the EB approach leads to compromise ($shrinkage$) posterior estimators between the individual (X_i) and group ($\bar{X}$) estimators.

One problem with the EB approach is that it clearly ignores the uncertainty in ϕ. Another problem with this approach is how to choose the estimators of the hyperparameters. Many options are possible, e.g. maximum likelihood, method of moments, unbiased estimators etc. and all will lead to slightly different solutions in general.

This is in contrast to the fully Bayesian approach which requires the definition of a hyperprior but avoids the necessity of selecting a given value of the hyperprior.

Fully hierarchical modeling

In order to implement a fully hierarchical model, we need to specify a hyperprior distribution $p(\boldsymbol{\phi})$. Then, we have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})\,d\phi \\ p(\boldsymbol{\phi}|\mathbf{x}) &\propto \int f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})\,d\boldsymbol{\theta}. \end{aligned}$$

In many cases, these integrals cannot be evaluated analytically. However, often a Gibbs sampling approach can be implemented by sampling from the conditional posterior distributions

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x},\boldsymbol{\phi}) &\propto f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\phi}) \\ p(\boldsymbol{\phi}|\mathbf{x},\boldsymbol{\theta}) &= p(\boldsymbol{\phi}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}) \end{aligned}$$

which often do have conjugate forms.

Example 78

Consider Example 74 and suppose initially that the values of ϕ and ψ are known and that we use a uniform distribution for μ. Then:

$$p(\mu, \boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{\phi}{2}\sum_i (y_i - \theta_i)^2 - \frac{\psi}{2}\sum_i (\theta_i - \mu)^2\right).$$

Integrating with respect to μ, we have $\boldsymbol{\theta}|\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{W})$ where

$$\mathbf{W}^{-1} = \frac{1}{\phi + \psi}\mathbf{I} + \frac{\psi}{n\phi(\phi + \psi)}\mathbf{J} \quad \text{and} \quad \mathbf{W}\mathbf{m} = \phi\mathbf{y}$$

and the posterior mean of $\boldsymbol{\theta}$ is given by

$$E[\boldsymbol{\theta}|\mathbf{y}] = \frac{\phi}{\phi + \psi}\mathbf{y} + \frac{\psi}{\psi + \phi}\mathbf{1}\bar{y}$$

which is a weighted average of the MLE and the global mean.

Suppose now that both ϕ and ψ are unknown and that we use the usual improper priors $p(\phi) \propto \frac{1}{\phi}$ and $p(\psi) \propto \frac{1}{\psi}$. Then it is easy to show that

$$\begin{aligned}
\boldsymbol{\theta}|\mathbf{y}, \mu, \phi, \psi &\sim \mathcal{N}\left(\frac{\phi\mathbf{y} + \psi\mu\mathbf{1}}{\phi + \psi}, \frac{1}{\phi + \psi}\mathbf{I}\right) \\
\mu|\mathbf{y}, \boldsymbol{\theta}, \phi, \psi &\sim \mathcal{N}\left(\bar{\theta}, \frac{1}{n\psi}\right) \\
\phi|\mathbf{y}, \boldsymbol{\theta}, \mu, \psi &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_i (y_i - \theta_i)^2}{2}\right) \\
\psi|\mathbf{y}, \boldsymbol{\theta}, \mu, \phi &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_i (\mu - \theta_i)^2}{2}\right)
\end{aligned}$$

and a Gibbs sampling algorithm could be set up. However, it is possible to demonstrate that the joint posterior distribution is *improper*.

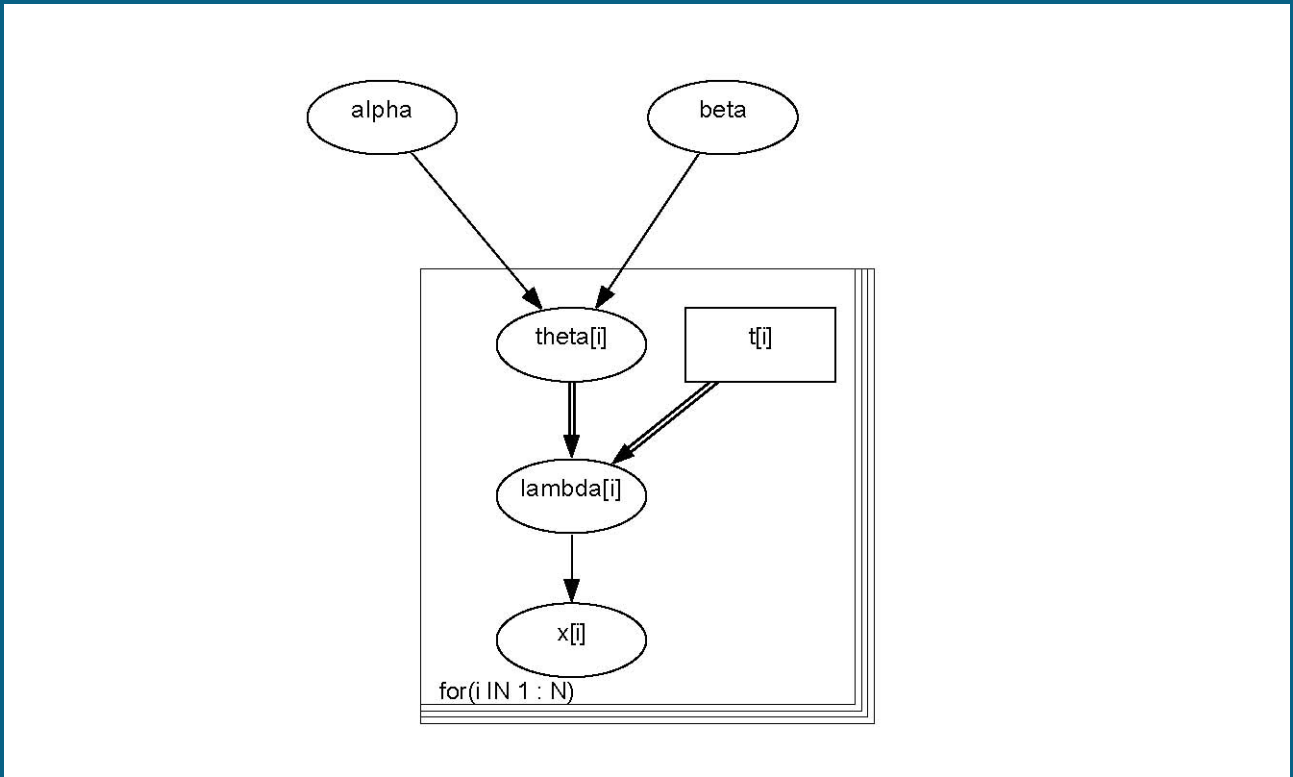
It is important to check the propriety of the posterior distribution when improper hyperprior distributions are used. An alternative (as in for example `Winbugs`) is to use proper but high variance hyperprior distributions.

Directed acyclic graph representations

Hierarchical models are often well represented by directed acyclic graphs or DAGs as used in `Winbugs`.

Example 79

A DAG representing the model and prior structure of Example 75 is as below.



Here we can see that X_i depends directly upon its rate λ_i which depends on t_i and θ_i through a logical relation ($\lambda_i = t_i\theta_i$).

In `Winbugs`, this can be converted into code for a Gibbs sampler.

```
model
{
    for (i in 1 : N) {
        theta[i] ~ dgamma(alpha, beta)
        lambda[i] <- theta[i] * t[i]
        x[i] ~ dpois(lambda[i])
    }
    alpha ~ dexp(1)
    beta ~ dgamma(0.1, 1.0)
}
```

George et al (1993) give the following table of pump failure data.

Pump	t_i	x_i
1	94.5	5
2	15.7	1
3	62.9	5
4	126	14
5	5.24	3
6	31.4	19
7	1.05	1
8	1.05	1
9	2.1	4
10	10.5	22

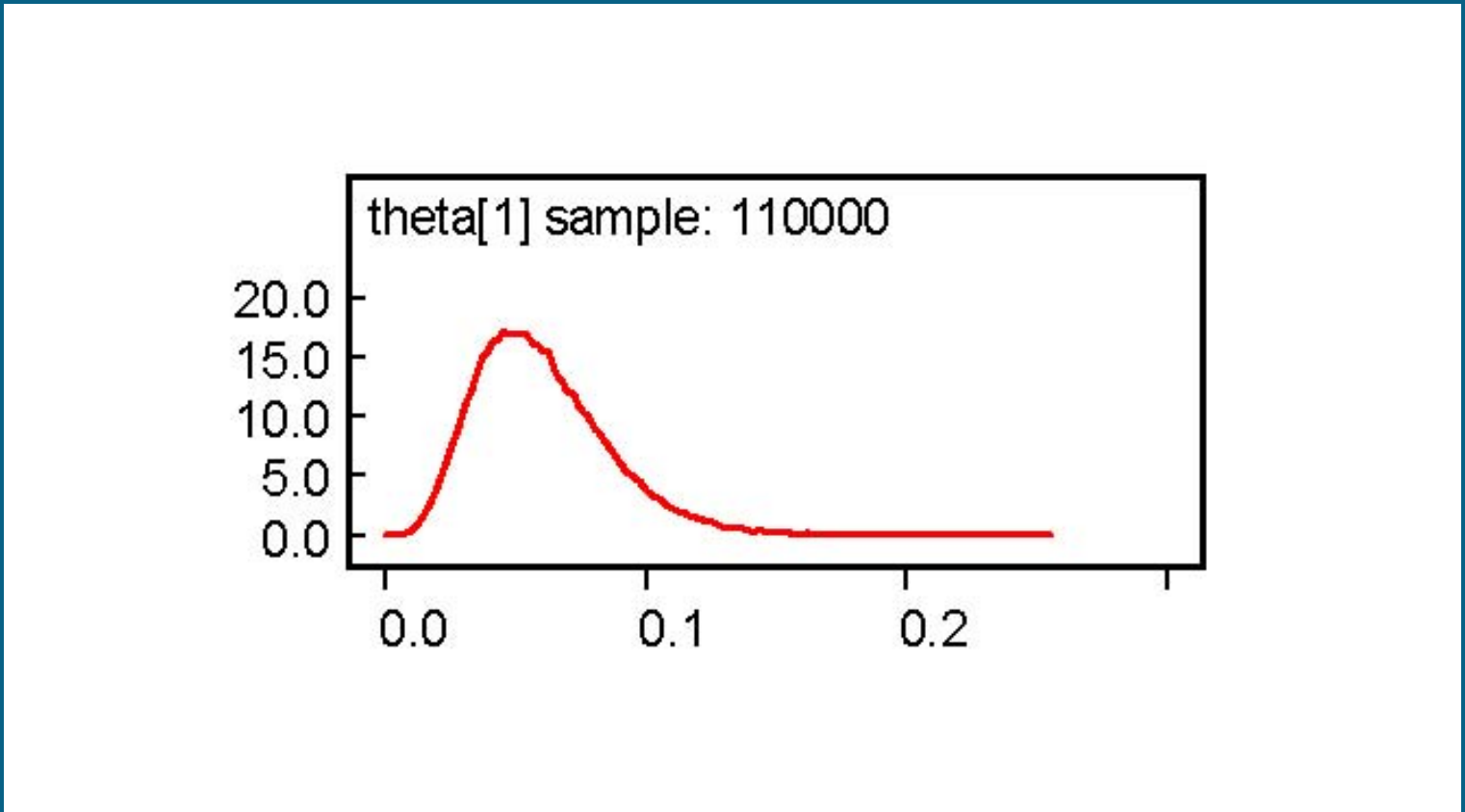
George et al (1993) assume a hierarchical prior distribution

$$\begin{aligned} \alpha &\sim \mathcal{E}(1) \\ \beta &\sim \mathcal{G}(0.1, 1) \end{aligned}$$

The following table gives the posterior means and variances of the different parameters estimated via `Winbugs`

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	0.7001	0.2699	0.004706	0.2851	0.6634	1.338	1001	10000
beta	0.929	0.5325	0.00978	0.1938	0.8315	2.205	1001	10000
theta[1]	0.0598	0.02542	2.68E-4	0.02128	0.05627	0.1195	1001	10000
theta[2]	0.1008	0.07855	8.177E-4	0.00838	0.08181	0.3023	1001	10000
theta[3]	0.08927	0.03759	3.702E-4	0.0316	0.08469	0.1762	1001	10000
theta[4]	0.116	0.03048	3.17E-4	0.06363	0.1132	0.1825	1001	10000
theta[5]	0.6056	0.315	0.003087	0.1529	0.5529	1.359	1001	10000
theta[6]	0.6105	0.1393	0.0014	0.3668	0.5996	0.9096	1001	10000
theta[7]	0.9025	0.7252	0.007937	0.07559	0.7167	2.751	1001	10000
theta[8]	0.8964	0.725	0.008262	0.07614	0.7098	2.785	1001	10000
theta[9]	1.59	0.7767	0.009004	0.4828	1.452	3.452	1001	10000
theta[10]	1.993	0.4251	0.004915	1.264	1.958	2.916	1001	10000

The last diagram shows a kernel density estimate of the posterior density of θ_1.



10. Time series and dynamic linear models

Objective

To introduce the Bayesian approach to the modeling and forecasting of time series.

Recommended reading

- West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*, (2'nd ed.). Springer.
- Pole, A.,West, M. and Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis.* Chapman and Hall.
- Bauwens, L., Lubrano, M. and Richard, J.F. (1999). *Bayesian inference in dynamic econometric models.* Oxford University Press.

Dynamic linear models

West

The first Bayesian approach to forecasting stems from Harrison and Stevens (1976) and is based on the *dynamic linear model.* For a full discussion, see West and Harrison (1997).

The general DLM

Definition 29
The general (univariate) dynamic linear model is

$$\begin{aligned} Y_t &= \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t \\ \boldsymbol{\theta}_t &= \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \end{aligned}$$

where ν_t and $\boldsymbol{\omega}_t$ are zero mean measurement errors and state innovations.

These models are linear *state space models*, where $x_t = \mathbf{F}_t^T \boldsymbol{\theta}_t$ represents the signal, $\boldsymbol{\theta}_t$ is the state vector, $\mathbf{F}_t$ is a regression vector and $\mathbf{G}_t$ is a state matrix. The usual features of a time series such as trend and seasonality can be modeled within this format.

In some cases, $\mathbf{F}$ and $\mathbf{G}$ are supposed independent of t. Then the model is a *time series DLM*. If V and $\mathbf{W}$ are also time independent then the DLM is *constant*.

Examples

Example 80
A slowly varying level model is

$$\begin{aligned} y_t &= \theta_t + \nu_t \\ \theta_t &= \theta_{t-1} + \omega_t \end{aligned}$$

The observations fluctuate around a mean which varies according to a random walk.

Example 81
A dynamic linear regression model is given by

$$\begin{aligned} y_t &= \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \end{aligned}$$

Bayesian analysis of DLM's

If the error terms, ν_t and $\boldsymbol{\omega}_t$ are normally distributed, with known variances, e.g. $\nu_t \sim \mathcal{N}(0, V_t)$, $\boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$, then a straightforward Bayesian analysis can be carried out.

Example 82

In Example 80, suppose that at time $t-1$, the current accumulated information is $D_{t-1} = \{y_1, y_2, \ldots, y_{t-1}\}$ and assume that the distribution for θ_{t-1} is $\theta_{t-1}|D_{t-1} \sim \mathcal{N}(m_{t-1}, C_{t-1})$. and that the error distributions are $\nu_t \sim \mathcal{N}(0, V_t)$ and $\omega_t \sim \mathcal{N}(0, W_t)$. Then, we have:

1. The prior distribution for θ_t is:

$$\begin{aligned} \theta_t|D_{t-1} &\sim \mathcal{N}(m_{t-1}, R_t) \quad \text{where} \\ R_t &= C_{t-1} + W_t \end{aligned}$$

2. The one step ahead predictive distribution for y_t is:

$$\begin{aligned} y_t|D_{t-1} &\sim \mathcal{N}(m_{t-1}, Q_t) \quad \text{where} \\ Q_t &= R_t + V_t \end{aligned}$$

3. The joint distribution of θ_t and y_t is

$$\begin{pmatrix} \theta_t \\ y_t \end{pmatrix} \Big| D_{t-1} \sim \mathcal{N}\left(\begin{pmatrix} m_{t-1} \\ m_{t-1} \end{pmatrix}, \begin{pmatrix} R_t & R_t \\ r_t & Q_t \end{pmatrix} \right)$$

4. The posterior distribution for θ_t given $D_t = \{D_{t-1}, y_t\}$ is

$$\begin{aligned} \theta_t|D_t &\sim \mathcal{N}(m_t, C_t) \quad \text{where} \\ m_t &= m_{t-1} + A_t e_t \\ A_t &= R_t/Q_t \\ e_t &= y_t - m_{t-1} \\ C_t &= R_t - A_t^2 Q_t. \end{aligned}$$

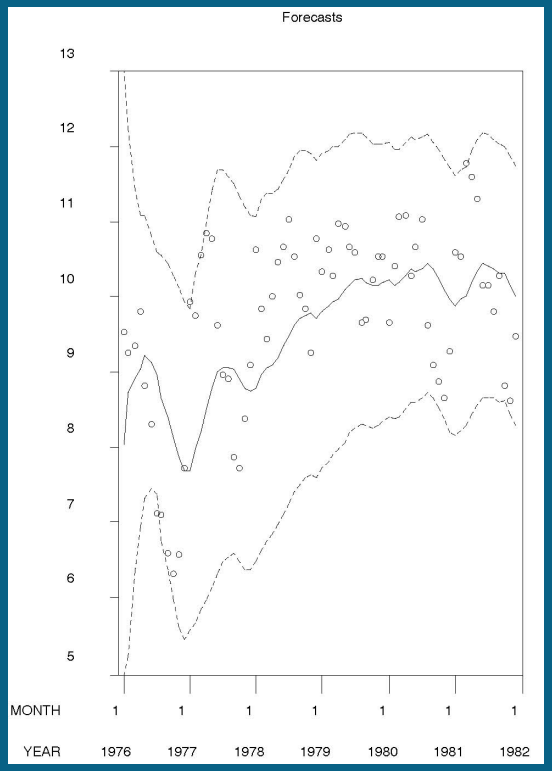
Proof and observations

Proof The first three steps of the proof are straightforward just by going through the observation and system equations. The posterior distribution follows from property iv) of the multivariate normal distribution as given in Definition 22. ∎

In the formula for the posterior mean, e_t is simply a prediction error term. This formula could also be rewritten as a weighted average in the usual way for normal models:

$$m_t = (1 - A_t)m_{t-1} + A_t y_t.$$

The following diagram illustrates the one step ahead predictions for the sales data from Pole et al (1994) assuming a model with constant observation and state error variances and a non-informative prior.



An interesting feature to note is that the predictive variance approaches a fixed constant for this model as the number of observed data increases. See West and Harrison (1997) for more details.

Example 83
In Example 81, suppose that we have $\nu_t \sim \mathcal{N}(0, V_t)$ and $\omega_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$ with distribution $\boldsymbol{\theta}_t | D_{t-1} \sim \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$. Then:

1. The prior distribution for $\boldsymbol{\theta}_t$ is:

$$\begin{aligned} \boldsymbol{\theta}_t | D_{t-1} &\sim \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{R}_t) \quad \text{where} \\ \mathbf{R}_t &= \mathbf{C}_{t-1} + \mathbf{W}_t. \end{aligned}$$

2. The one step ahead predictive distribution for y_t is:

$$\begin{aligned} y_t | D_{t-1} &\sim \mathcal{N}(f_t, Q_t) \quad \text{where} \\ f_t &= \mathbf{F}_t^T \mathbf{m}_{t-1} \\ Q_t &= \mathbf{F}_t^T \mathbf{R}_t \mathbf{F}_t + V_t \end{aligned}$$

3. The joint distribution of $\boldsymbol{\theta}_t$ and y_t is

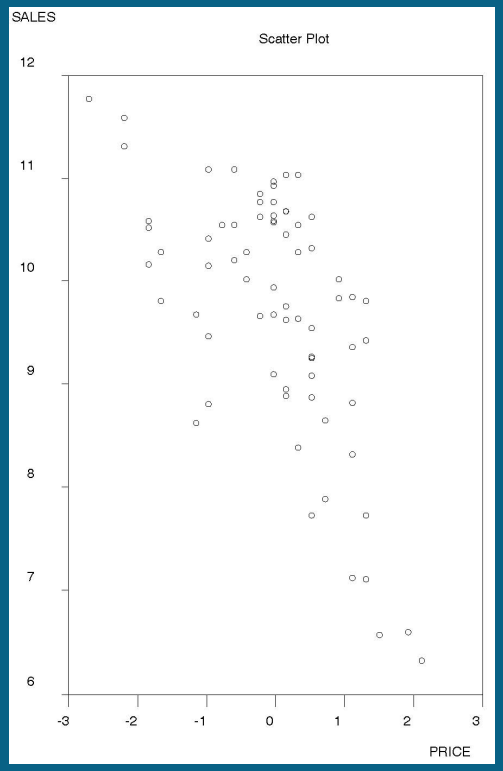
$$\begin{pmatrix} \boldsymbol{\theta}_t \\ y_t \end{pmatrix} \Big| D_{t-1} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}_{t-1} \\ f_t \end{pmatrix}, \begin{pmatrix} \mathbf{R}_t & \mathbf{F}_t^T \mathbf{R}_t \\ \mathbf{R}_t \mathbf{F}_t & Q_t \end{pmatrix} \right)$$

4. The posterior distribution for $\boldsymbol{\theta}_t$ given $D_t = \{D_{t-1}, y_t\}$ is

$$\begin{aligned} \boldsymbol{\theta}_t | D_t &\sim \mathcal{N}(\mathbf{m}_t, \mathbf{C}_t) \quad \text{where} \\ \mathbf{m}_t &= \mathbf{m}_{t-1} + \mathbf{A}_t e_t \\ \mathbf{C}_t &= \mathbf{R}_t - \mathbf{A}_t \mathbf{A}_t^T Q_t \\ \mathbf{A}_t &= \mathbf{R}_t \mathbf{F}_t Q_t^{-1} \\ e_t &= y_t - f_t. \end{aligned}$$

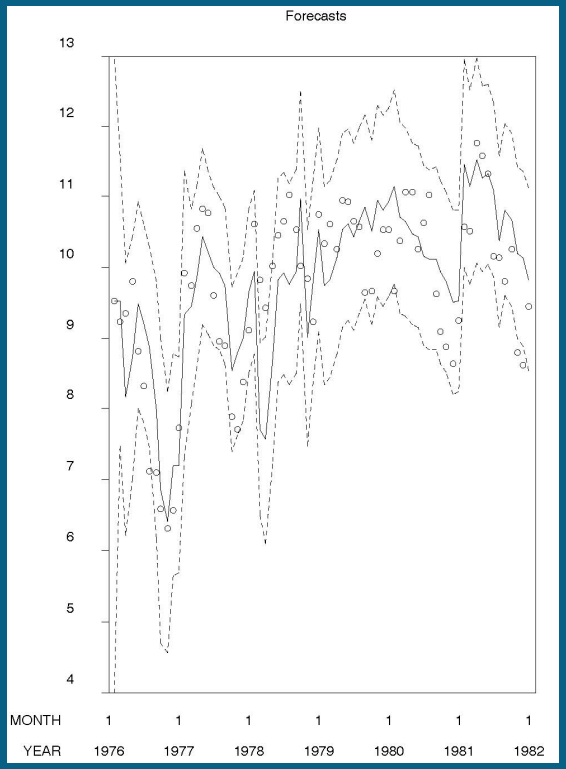
Proof Exercise. ∎

The following plot shows sales against price.



Thus, a dynamic, simple linear regression model would seem appropriate.

The following diagram, assuming a constant variance model as earlier illustrates the improved fit of this model.



The general theorem for DLM's

Theorem 43
For the general, univariate DLM,

$$\begin{aligned} Y_t &= \mathbf{F}_t^T \boldsymbol{\theta}_t + \nu_t \\ \boldsymbol{\theta}_t &= \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \end{aligned}$$

where $\nu_t \sim \mathcal{N}(0, V_t)$ and $\omega_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$, assuming the prior distribution $\boldsymbol{\theta}_{t-1}|D_{t-1} \sim \mathcal{N}(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$, we have

1. Prior distribution for $\boldsymbol{\theta}_t$:

$$\begin{aligned} \boldsymbol{\theta}_t|D_{t-1} &\sim \mathcal{N}(\mathbf{a}_t, \mathbf{R}_t) \quad \text{where} \\ \mathbf{a}_t &= \mathbf{G}_t \mathbf{m}_{t-1} \\ \mathbf{R}_t &= \mathbf{G}_t \mathbf{C}_{t-1} \mathbf{G}_t^T + \mathbf{W}_t \end{aligned}$$

2. One step ahead prediction:

$$\begin{aligned} y_t|D_{t-1} &\sim \mathcal{N}(f_t, Q_t) \quad \text{where} \\ f_t &= \mathbf{F}_t^T \mathbf{a}_t \\ Q_t &= \mathbf{F}_t^T \mathbf{R}_t \mathbf{F}_t + V_t. \end{aligned}$$

3. Posterior distribution for $\boldsymbol{\theta}_t|D_t$:

$$\begin{aligned} \boldsymbol{\theta}_t|D_t &\sim \mathcal{N}(\mathbf{m}_t, \mathbf{C}_t) \quad \text{where} \\ \mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t e_t \\ \mathbf{C}_t &= \mathbf{R}_t \mathbf{R}_t^T Q_t \\ \mathbf{A}_t &= \mathbf{R}_t \mathbf{F}_t Q_t^{-1} \\ e_t &= y_t - f_t. \end{aligned}$$

Proof Exercise. ∎

DLM's and the Kalman filter

The updating equations in the general theorem are essentially those used in the Kalman filter developed in Kalman (1960) and Kalman and Bucy (1961). For more details, see

http://en.wikipedia.org/wiki/Kalman_filter

Superposition of models

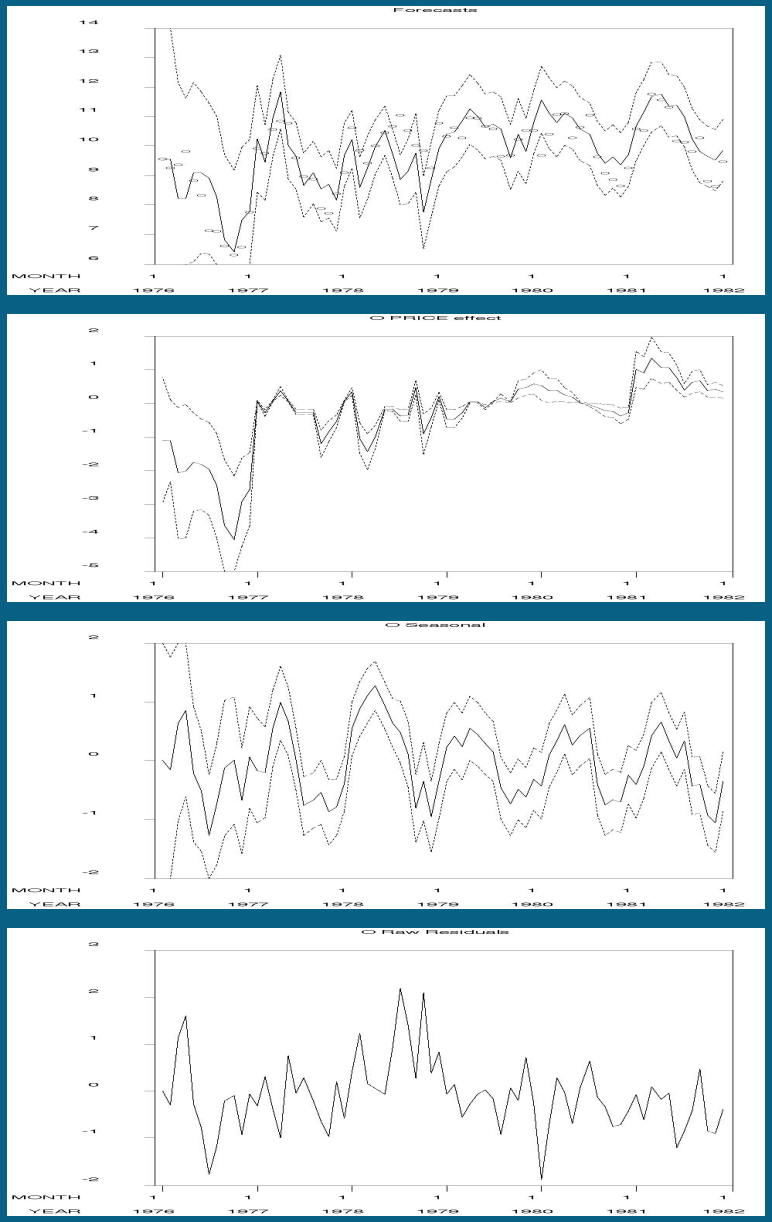
Many time series exhibit various different components. For example, as well as the regression component we have already fitted, it may well be that the sales series exhibits a seasonal component. In such cases, we may often wish to combine these components in a single model. In such cases, we may write

$$\begin{aligned} y_t &= y_{1t} + \ldots + y_{kt} + \nu_t \quad \text{where} \\ y_{jt} &= \mathbf{F}_{jt}^T \boldsymbol{\theta}_{jt} \quad \text{and} \\ \boldsymbol{\theta}_{jt} &= \mathbf{G}_{jt} \boldsymbol{\theta}_{j,t-1} + \boldsymbol{\omega}_{jt} \quad \text{for } j = 1, \ldots, k. \end{aligned}$$

This leads to a combined model

$$\begin{aligned} y_t &= \mathbf{F}_t^T \boldsymbol{\theta} + \nu_t \\ \boldsymbol{\theta}_t &= \mathbf{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \quad \text{where} \end{aligned}$$

$$\mathbf{F}_t = \begin{pmatrix} \mathbf{F}_{1t} \\ \vdots \\ \mathbf{F}_{kt} \end{pmatrix}, \quad \mathbf{G}_t = \begin{pmatrix} \mathbf{G}_{1t} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{G}_{kt} \end{pmatrix}.$$


Forecasts
PRICE effect
Seasonal
Raw Residuals
MONTH
YEAR
1976
1977
1978
1979
1980
1981
1982

Discount factors

Thus far, we have not considered how to model the uncertainty in the unknown variances. It is possible to model the uncertainty in the observation variances analytically in the usual way (via inverse gamma priors). However, the treatment of the system variances is more complex. In this case, *discount factors* can be used.

The idea is based on information discounting. As information ages, it becomes less useful and so its value should diminish. Thus, in our problem, with system equation

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$$

then given that $V[\boldsymbol{\theta}_{t-1}|D_{t-1}] = \mathbf{C}_{t-1}$, we have

$$\mathbf{R}_t = V[\boldsymbol{\theta}_t|D_{t-1}] = \mathbf{P}_t + \mathbf{W}_t$$

where

$$\mathbf{P}_t = V[\mathbf{G}_t\boldsymbol{\theta}_{t-1}|D_{t-1}] = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t^T \quad \text{and} \quad \mathbf{W}_t = \mathbf{R}_t - \mathbf{P}_t.$$

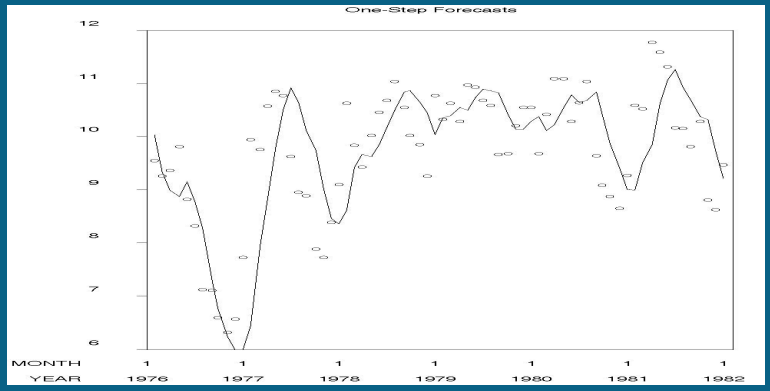
If we define δ such that $\mathbf{R}_t = \mathbf{P}_t/\delta$, then we can interpret δ as the percentage of information that passes from time $t-1$ to time t and in this case,

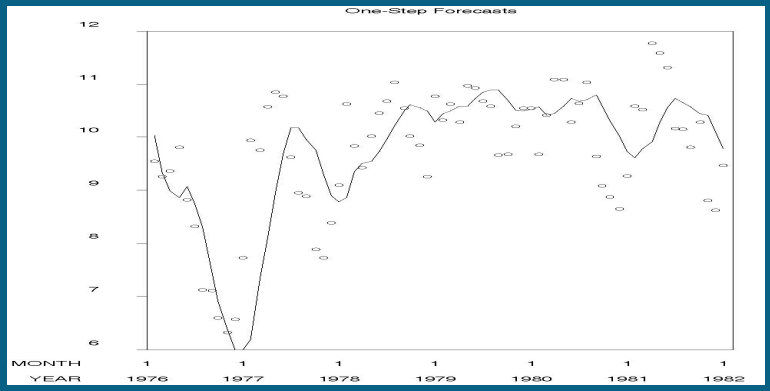
$$\mathbf{W}_t = \mathbf{P}_t \left(\delta^{-1} - 1\right).$$

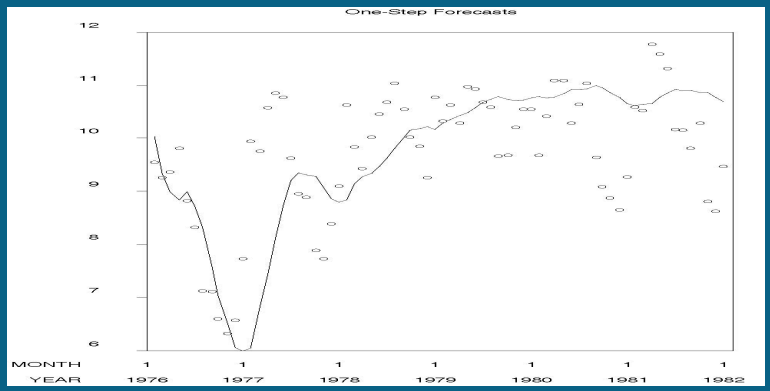
Typical values for systems without abrupt changes are usually around $\delta = 0.9$. Small values of δ (below 0.8) imply large levels of uncertainty and lead to predictions with very wide bounds.

High values represent more smoothly changing systems, and in the limit, when $\delta = 1$, we have a static system with no information loss.

The following diagrams show the effects of fitting a trend model with discount factors 0.8, 0.9 and 1 to the sales data. We can see that the higher the discount factor, the higher the degree of smoothing.


One-Step Forecasts
12
11
10
9
8
7
6
MONTH 1 1 1 1 1 1 1
YEAR 1976 1977 1978 1979 1980 1981 1982


One-Step Forecasts
12
11
10
9
8
7
6
MONTH 1 1 1 1 1 1 1
YEAR 1976 1977 1978 1979 1980 1981 1982


One-Step Forecasts
12
11
10
9
8
7
6
MONTH 1 1 1 1 1 1 1
YEAR 1976 1977 1978 1979 1980 1981 1982

The forward filtering backward sampling algorithm

This algorithm, developed in Carter and Kohn (1994) and Frühwirth-Schnatter (1994) allows for the implementation of an MCMC approach to DLM's.

The forward filtering step is the standard normal linear analysis to give $p(\boldsymbol{\theta}_t|D_t)$ at each t, for $t = 1, \ldots, n$.

The backward sampling step uses the Markov property and samples $\boldsymbol{\theta}_n^\star$ from $p(\boldsymbol{\theta}_n|D_n)$ and then, for $t = 1, \ldots, n-1$, samples $\boldsymbol{\theta}_t^\star$ from $p(\boldsymbol{\theta}_t|D_t, \boldsymbol{\theta}_{t+1}^\star)$. Thus, a sample from the posterior parameter structure is generated.

Example: the $AR(p)$ model with time varying coefficients

Example 84

The $AR(p)$ model with time varying coefficients takes the form

$$\begin{aligned} y_t &= \theta_{0t} + \theta_{1t} + \ldots + \theta_{pt} y_{t-p} + \nu_t \\ \theta_{it} &= \theta_{i,t-1} + \omega_{it} \end{aligned}$$

where we shall assume that the error terms are independent normals:

$$\nu_{it} \sim \mathcal{N}(0, V) \quad \text{and} \quad \omega_{it} \sim \mathcal{N}(0, \lambda_i V).$$

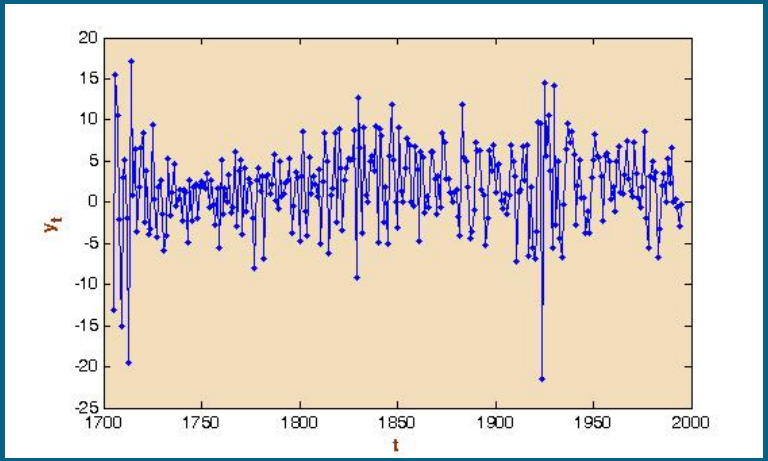
Then this model can be expressed in state space form by setting

$$\begin{aligned} \boldsymbol{\theta}_t &= (\theta_{0t}, \ldots, \theta_{pt})^T \\ \mathbf{F} &= (1, y_{t-1}, \ldots, y_{t-p})^T \\ \mathbf{G} &= \mathbf{I}_{p+1} \\ \mathbf{W} &= V \mathrm{diag}(\boldsymbol{\lambda}) \end{aligned}$$

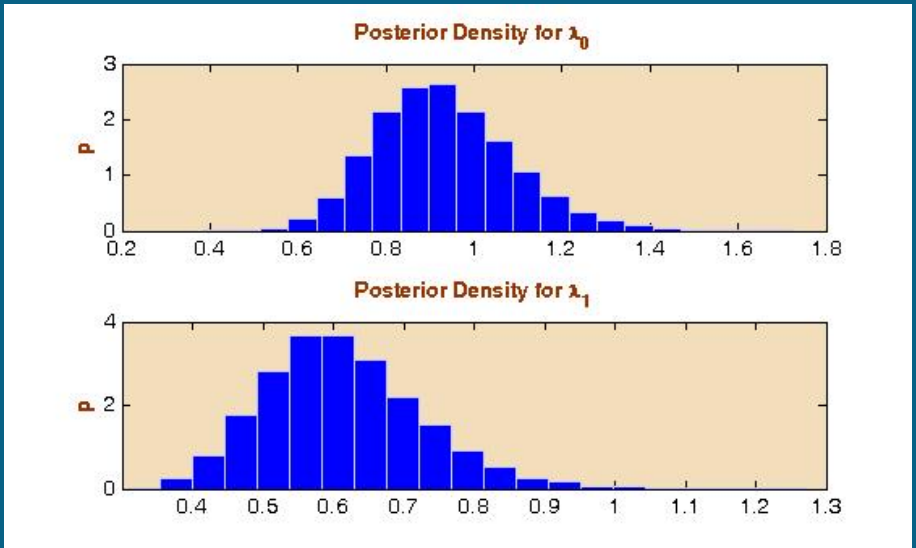
Here $\mathrm{diag}(\boldsymbol{\lambda})$ represents a diagonal matrix with ii'th entry equal to λ_i, for $i = 1, \ldots, p+1$.

Now, given gamma priors for V and for λ_i^{-1} and a normal prior for $\boldsymbol{\theta}_0$, then it is clear that the relevant posterior distributions are all conditionally conjugate.

Koop (2003) examines data on the annual percentage change in UK industrial production from 1701 to 1992.



Koop (2003) assumes a time varying coefficient AR(1) model and uses proper but relatively uninformative prior distributions. The posterior distributions of λ_0 and λ_1 estimated from running the Gibbs sampler are as follows.



The posterior means and deviations of both λ_0 and λ_1 suggest that there is quite high stochastic variation in both θ_{0t} and θ_{1t}.

Software for fitting DLM's

Two general software packages are available.

- `BATS`. Pole et al (1994). This (somewhat out of date) package can be used to perform basic analyses and is available from:

  `http://www.stat.duke.edu/~mw/bats.html`

- `dlm`. Petris (2006). This is a recently produced R package for fitting DLM's, including ARMA models etc. available from

  `http://cran.r-project.org/src/contrib/Descriptions/dlm.html`

Other work in time series

- ARMA and ARIMA models. Marriot and Newbold (1998).
- Non linear and non normal state space models. Carlin et al (1992).
- Latent structure models. Aguilar et al (1998).
- Stochastic volatility models. Jacquier et al (1994).
- GARCH and other econometric models. Bauwens y Lubrano (1998), Bauwens et al (2000).
- Wavelets. Nason et al (1999).

11. Other topics

Objective

Introduce the basic ideas of robust Bayesian analysis and nonparametric Bayesian methods.

Recommended reading

- Berger, J. (1994) An overview of robust Bayesian analysis (with discussion). *Test*, **3**, 5–124.

- Ríos Insua, D. and Ruggeri, F. (eds.) (2000). *Robust Bayesian Analysis*. Springer Verlag.

Robustness

In any Bayesian analysis, especially when expert priors have been solicited, it is important to assess the sensitivity of the results to the election of the prior distribution $p(\boldsymbol{\theta})$.

Sensitivity to the loss function and likelihood function are also considered in Dey and Micheas (2000), Kadane et al (2000) and Shyamalkumar (2000).

An informal sensitivity analysis considers robustness to the use of various different prior distributions.

Example

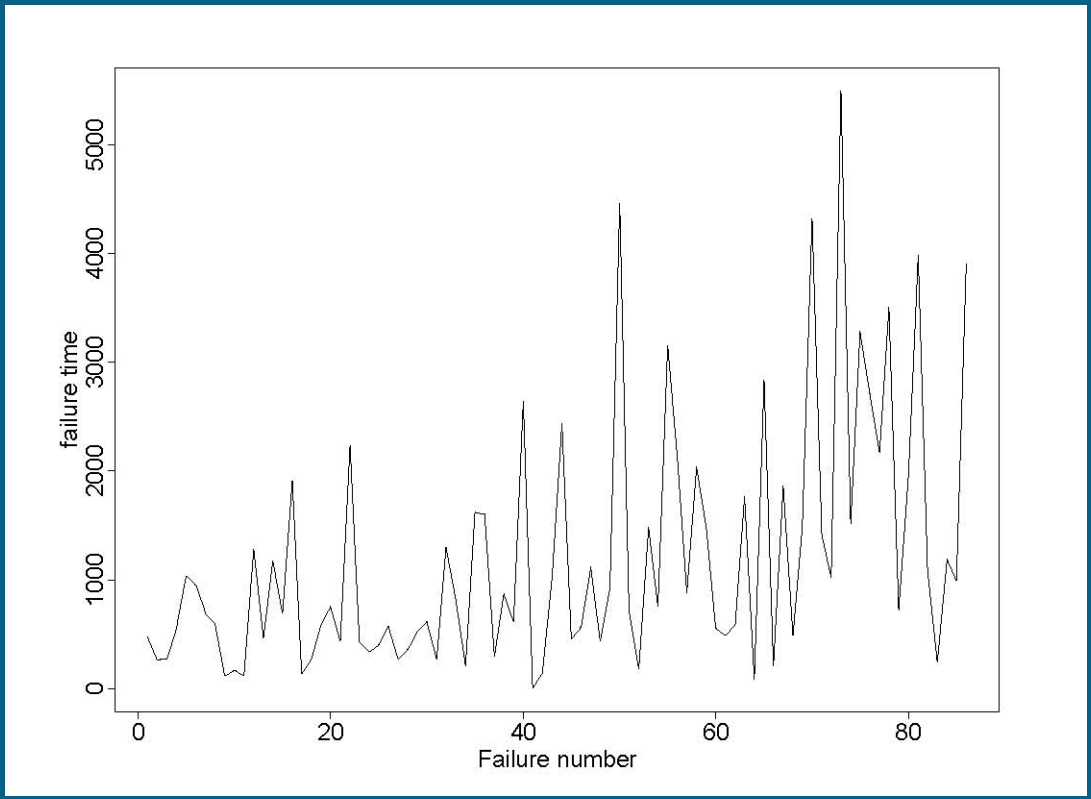
Example 85
Wilson and Wiper (2000) analyzed the Jelinski Moranda (1972) model for software reliability.

Let $T_1, T_2, \ldots$ be the times between successive software failures. Then the Jelinski Moranda model assumes that

$$T_i | N, \phi \sim \mathcal{E}\left((N - i + 1)\phi\right)$$

so that initially, the program contains N faults, each of which has the same size or importance, ϕ, and then after each failure, the fault causing the failure is identified and removed from the program.

The first $m = 86$ inter failure times for a program were observed as in the following diagram.



The objective is to predict the next failure time and the number of bugs left in the program.

The likelihood function is

$$l(N, \phi|\mathbf{t}) \propto \frac{N!}{(N-m)!}\phi^m \exp\left(-\left[(N+1)m\bar{t} - \sum_{i=1}^{m} it_i\right]\phi\right)$$

and semi-conjugate priors are

$$\begin{aligned} N &\sim \mathcal{P}(\lambda) \quad \text{where we assume that } \lambda = 100 \\ \phi &\sim \mathcal{G}(\alpha, \beta) \quad \text{where } \alpha = 1 \text{ and } \beta = .0001 \end{aligned}$$

Given these priors, it can be shown that $E[N|\text{data}] \approx 104$ (MLE $= 106$) and the posterior median of the distribution of the time to next failure is 2440×10^{-2} seconds (MLE $= 2177$).

Assume that we contaminate the prior distribution with a long tailed distribution. We shall consider the class of prior distributions

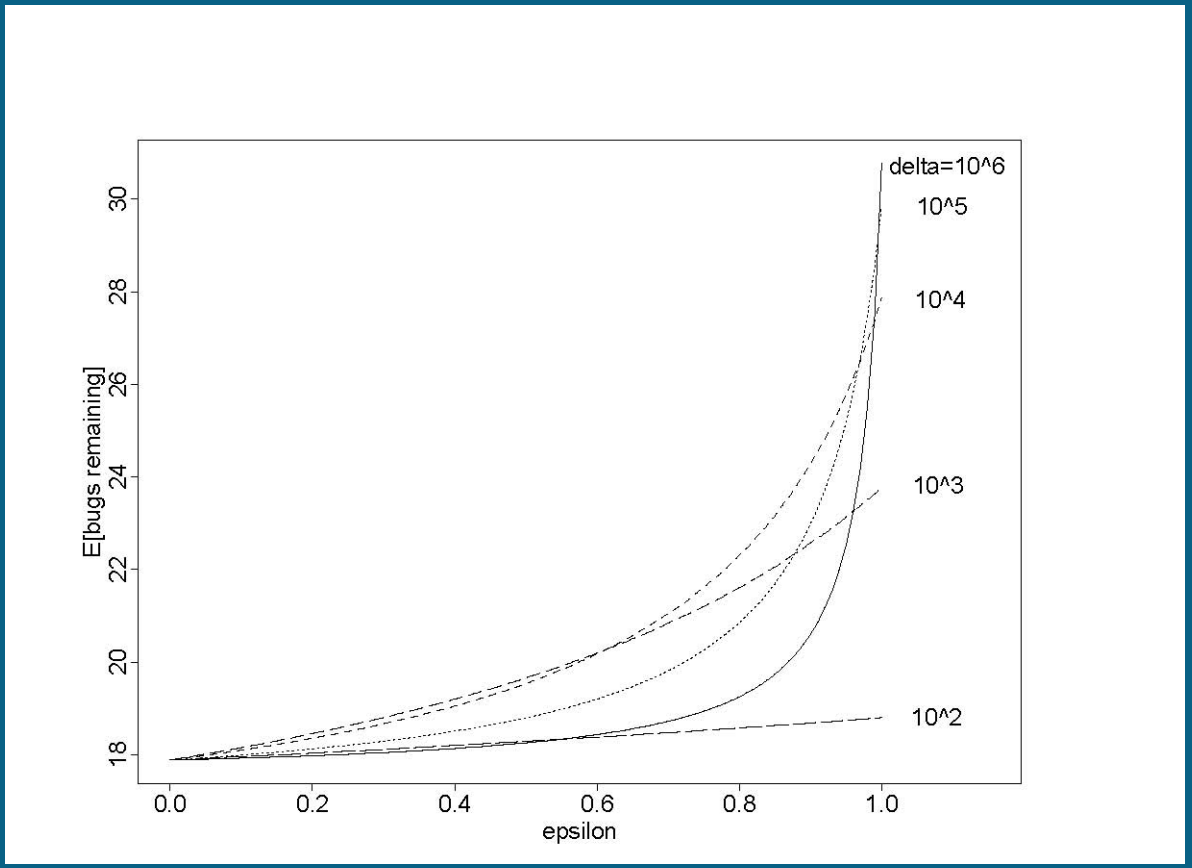
$$\Gamma = \{(1-\epsilon)P(N) + \epsilon Q(N)\}$$

where $P(N)$ represents the Poisson density and

$$Q(N) \propto \left((N - \lambda + \frac{1}{2})^2 + \delta\right)^{-1}.$$

Q is a density with the same mode as the Poisson but with no mean.

The following diagram illustrates the effects of the contamination on the posterior mean of the number of remaining faults for different values of ϵ and δ.



For $\epsilon = 0.2$ the posterior median of the time to next failure varies between 2410 and 2440 for $100 \leq \delta \leq 1000000$ but when $\epsilon = 1$, the median is 1960 in the worst case, $\delta = 1000000$. We can conclude that the results are relatively insensitive to small changes in the prior.

Global sensitivity analysis

In the global approach, the prior is included within a wider class of distributions, Γ, and the sensitivity of some function of interest such as the posterior mean is examined. If there are large differences between the maximum and minimum estimates over the prior class then the inference is sensitive.

Possible classes are:

- ϵ-contamination classes

$$\Gamma = \{\pi : \ \pi(\theta) = (1-\epsilon)P(\theta) + \epsilon Q(\theta), \ g \in \mathcal{Q}\}$$

  where $\mathcal{Q}$ is a general class of contaminating distributions, e.g. unimodal distributions.

- Generalized moment classes: that is all distributions with a given set of specified moments or quantiles.

- Classes of density bands:

$$\Gamma = \{\pi : \ L(\theta) < \pi(\theta) < U(\theta)\}$$

for example, $L(\theta) = (1 - \epsilon) f(\theta)$ and $U(\theta) = (1 + \epsilon) f(\theta)$.

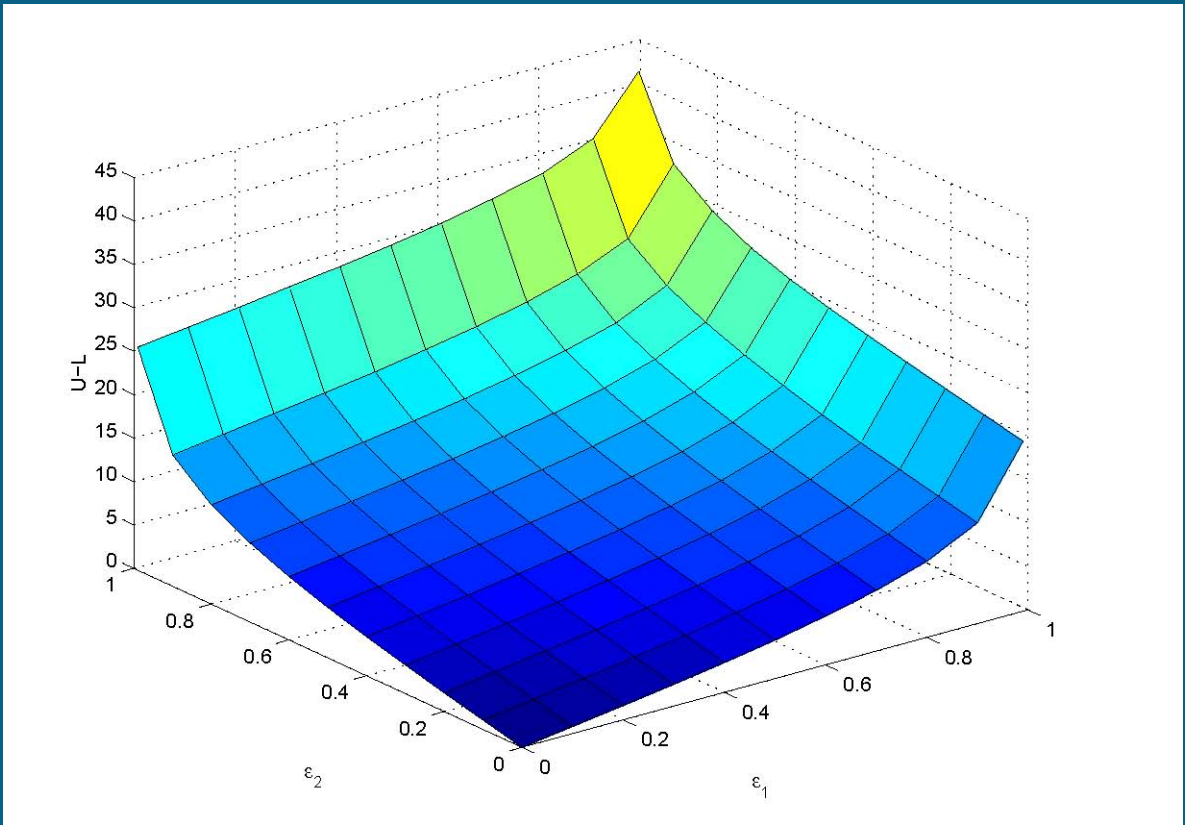
Example 86

Returning to the previous example, suppose that we wish to maintain the assumption of prior independence between N and ϕ. Then, we can define the class

$$\Gamma = \{\pi : \ \pi(N, \phi) = \pi_1(N)\pi_2(\phi)\}$$

where $(1 - \epsilon_1)P(N) < \pi_1(N) < (1 + \epsilon_1)P(N)$ and $(1 - \epsilon_2)P(\phi) < \pi_2(\phi) < (1 + \epsilon_2)P(\phi)$ and $P(N)$ and $P(\phi)$ are the Poisson and gamma priors we assumed earlier.

The diagram illustrates the differences between the upper and lower limits in the value of the posterior mean of N for $0 < \epsilon_1, \epsilon_2 < 1$.



The inference is more sensitive to contaminations of the prior of N than to contaminations in the prior for ϕ although it is still quite robust to small contaminations.

Problems with the global robustness approach

The main difficulty with this approach is that it is not always clear how to elect a reasonable contamination class. Most standard classes are too large and include unreasonable choices of prior.

A second problem is that the calculation of the minima and maxima of the quantity of interest is often very complex except for a very few, relatively simple contamination classes where analytic results are known.

Local robustness

An alternative approach is local robustness. In this case, influence measures based on, for example, the norm of the Frechet derivative of the posterior distribution relative to the prior are used to reflect the sensitivity to the the prior distribution. See e.g. Gustafson and Wassermann (1995) for more details.

Bayesian nonparametrics

Suppose that $X|f \sim f$ and that given a sample, $\mathbf{x}$, we wish to carry out inference about f.

In order to do this from a Bayesian standpoint, it is necessary to define a prior distribution over the (infinite dimensional) space of distributions. The simplest and most studied class of prior distributions is Dirichlet process priors.

The Dirichlet process

The definition of the Dirichlet process is a generalization of the Dirichlet distribution. It was first considered by Ferguson (1973).

Definition 30
Suppose that F is a random probability measure and that t F_0 is a (known) distribution function and α a scalar parameter. For any finite partition, $\{C_1, \ldots, C_r\}$, of the probability space, the Dirichlet process prior distribution for F, with parameters α and F_0 assigns the distribution

$$\{F(C_1), \ldots, F(C_r)\} \sim \mathcal{D}\left(\alpha F_0(C_1), \ldots, \alpha F_0(C_r)\right).$$

In this case, we write $F \sim \mathcal{DP}(\alpha, F_0)$.

It is straightforward to show that the Dirichlet process prior is conjugate.

Theorem 44
If $\{X\}_i$ is a sequence of exchangeable random variables with $X_i|F \sim F$ and $F \sim \mathcal{DP}(\alpha, F_0)$, then:

- the marginal distribution of X_i is F_0.

- the conditional distribution of F given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, is also a Dirichlet process such that

$$\{F(C_1), \ldots, F(C_r)\}|\mathbf{x} \sim \mathcal{D}\left(\alpha F_0(C_1) + \sum_{i=1}^{n} I_{C_1}(x_i), \ldots, \alpha F_0(C_r) + \sum_{i=1}^{n} I_{C_r}(x_i)\right)$$

that is $F|\mathbf{x} \sim \mathcal{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n}F_0 + \frac{n}{\alpha+n}\hat{F}\right)$ where $\hat{F}$ is the empirical c.d.f.

Proof Firstly, the marginal c.d.f. of X is

$$\begin{aligned} P(X \leq x) &= \int P(X \leq x|F)p(F)\,dF \\ &= \int F(x)p(F)\,dF \\ &= \frac{\alpha F_0(x)}{\alpha} = F_0(x) \end{aligned}$$

Secondly, we have

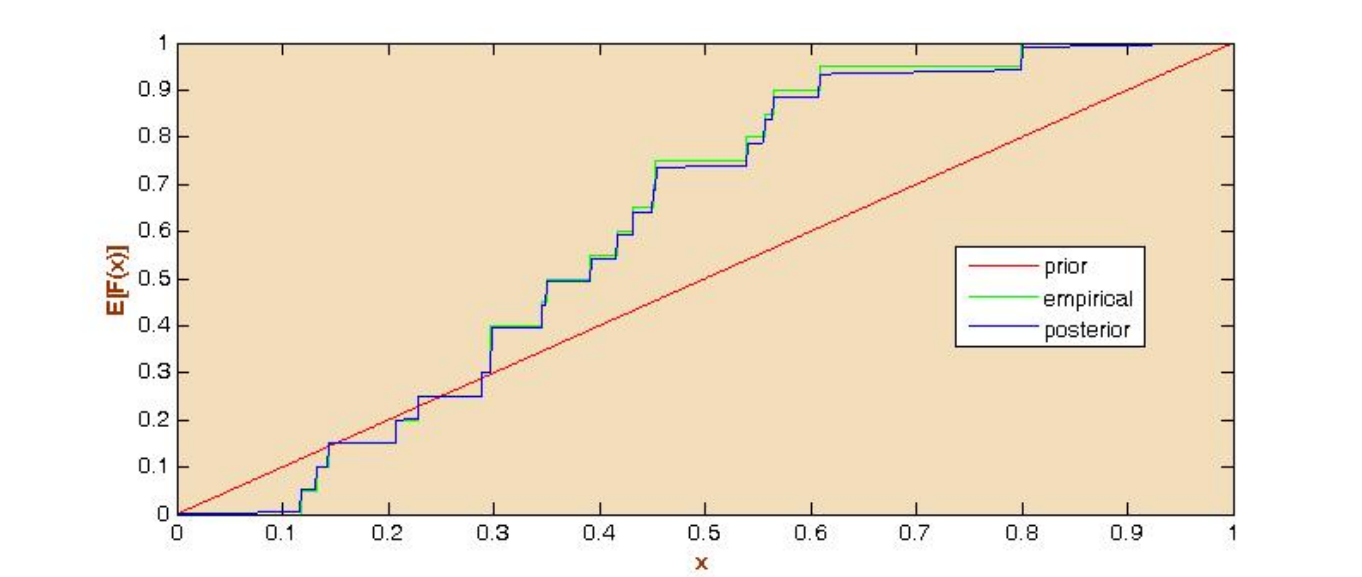
$$\begin{aligned} P(F(C_1), \ldots, F(C_r)|\mathbf{x}) &\propto P(F(C_1), \ldots, F(C_r))f(\mathbf{x}|F(C_1), \ldots, F(C_r)) \\ &\propto \prod_{j=1}^{r} F(C_j)^{\alpha F_0(C_i)+\sum_{i=1}^{n} I_{C_j}(x_i)} \end{aligned}$$

which is another Dirichlet distribution and proves the result. ∎

When n increases, the predictive distribution function of X_{n+1} approaches the empirical c.d.f.

Example 87

20 data were generated from a beta distribution, $\mathcal{B}(4, 6)$. A Dirichlet process prior with $\alpha = 1$ and $F_0(x) = x$, for $0 < x < 1$, i.e. a uniform distribution was assumed. The following diagram shows the predictive and empirical distribution functions.



Mixtures of Dirichlet processes

An important theoretical disadvantage of the Dirichlet process is that it can be shown that it assigns probability one to discrete probability measures, see Blackwell (1973). One way of getting around this is to use continuous mixtures of Dirichlet processes as developed in Antoniak (1974).

This leads to a hierarchical model

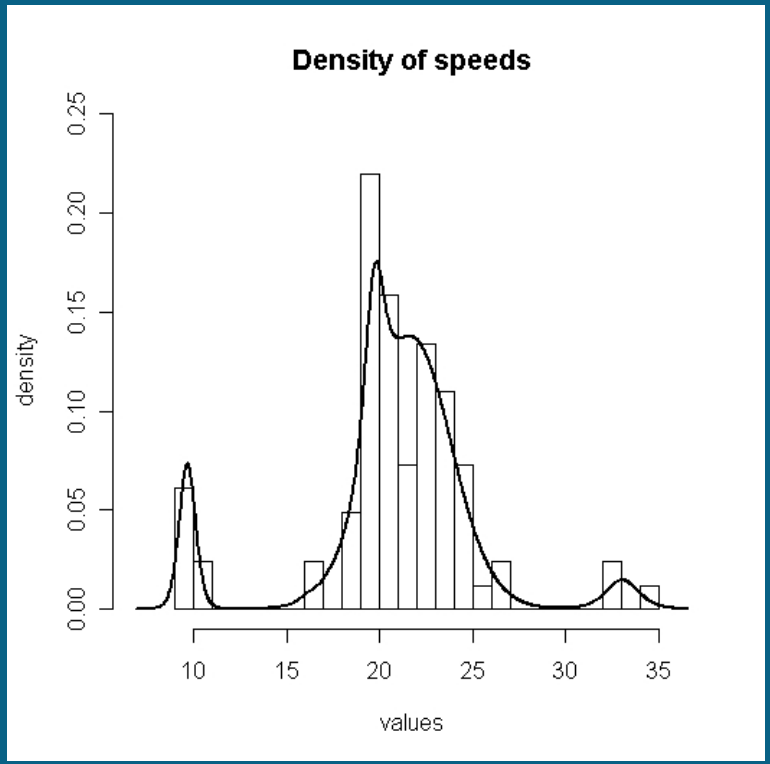
$$\begin{aligned} X_i|\theta_i &\sim f(x|\theta_i) \\ \theta_i|P &\sim P(\theta) \\ P &\sim \mathcal{DP}(\alpha, P_0) \end{aligned}$$

This model can be interpreted as a countably infinite, mixture model, i.e. as the limit of a finite mixture model, $f(x) = \sum_{i=1}^{k} w_i f(x|\theta_i)$, as the number of terms, k, goes to infinity.

One problem is how to choose the conditional density $f(x|\theta)$. Most applications have chosen to use a normal density as a flexible option. See e.g. McEachern and Muller (1998) or Neal (2000). In this case, inference is then carried out using MCMC techniques.

Example 88

The following diagram shows a fit of the well known galaxy data using the DP mixture model.



Nonparametric regression

A nonparametric regression model can be expressed as

$$y_i = f(x_i) + \epsilon_i$$

for $i = 1, \ldots, n$ where the function f is unknown. A number of classical estimation techniques are available for fitting such models, e.g. splines, neural networks or SVM's.

Bayesian penalized regression splines are implemented in Winbugs by e.g. Crainiceanu et al (2007) and a review of Bayesian neural nets is given by Lee (2004). The Bayesian equivalent of an SVM is the Gaussian process.

The Gaussian process

A Gaussian process defines a distribution over functions, f, where f is a function mapping some input space, $\mathcal{X}$ into $\mathbb{R}$. We shall call this distribution $P(f)$.

Let $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ be an n dimensional vector of function points evaluated at $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then $\mathbf{f}$ is a random variable.

Now, $P(f)$ is a Gaussian process if for any finite set, $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subset \mathcal{X}$ then the marginal distribution $P(\mathbf{f})$ over that subset has a multivariate Gaussian distribution.

Gaussian processes are characterized by a mean value function $\mu(\mathbf{x})$ and a covariance function $c(\mathbf{x}, \mathbf{x}')$, so that

$$\begin{aligned} P(f(\mathbf{x}), f(\mathbf{x}') &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{where} \\ \boldsymbol{\mu} &= \begin{pmatrix} \mu(\mathbf{x}) \\ \mu(\mathbf{x}') \end{pmatrix} \\ \boldsymbol{\Sigma} &= \begin{pmatrix} c(\mathbf{x}, \mathbf{x}) & c(\mathbf{x}, \mathbf{x}') \\ c(\mathbf{x}', \mathbf{x}) & c(\mathbf{x}', \mathbf{x}') \end{pmatrix} \end{aligned}$$

and similarly. Various forms for the covariance function have been considered, e.g.

$$c(x_i, x_j) = \nu_0 \exp\left(-\frac{|x_i - x_j|^\alpha}{\lambda}\right) + \nu_1 + \nu_2 \delta_{ij}$$

Some software for Gaussian process regression is available. See e.g.

`http://www.gaussianprocess.org/gpml/code/matlab/doc/regression.html`

Example