Bayesian Statistics
Script for the course in spring 2014

Hansruedi Künsch
Seminar für Statistik, ETH Zurich

Version of August 6, 2014

Contents

Remarks: This current version contains all the material that I managed to cover during the lecture. The last Section (4.4) goes beyond what we discussed during the course. Compared with the version of May 20, only minor typos have been corrected and at a few places some intermediate steps in the derivation have been spelled out. I will continuously correct mistakes and make smaller changes that should help to understand the material better. Please inform me about mistakes that you find and tell me about parts that are particularly obscure any time by sending an email to `kuensch@stat.math.ethz.ch`.

Originally I planned to have a fifth Chapter on Nonparametric Bayes Methods. I might add such a chapter later in fall or next year, but I don't want to commit myself now.

This script uses material from the two books "The Bayesian choice", 2nd edition, by C. Robert (Springer 2007) and A. Gelman et al., Bayesian Data Analysis, 3rd edition, Chapman & Hall (2013), and from the website by Michael Jordan for his course "Bayesian modeling and inference", `http://www.cs.berkeley.edu/~jordan/courses/260-spring10/`, but the organization and selection of topics and results is my own.

Chapter 1

Basic concepts

1.1 Interpretations of probability

Jakob Bernoulli wrote “Probability is the degree of certainty which is to the certainty as a part is to a whole”. There are however two somewhat different kinds of (un)certainty, called “aleatoric uncertainty” and “epistemic uncertainty”. Aleatoric uncertainty (from Latin “alea” = dice) is the uncertainty about outcomes of repeatable events like throwing a dice. In such cases, probabilities can be understood as mathematical idealizations of long-run relative frequencies, which we call the frequentist interpretation of probability. Epistemic uncertainty (from Greek “episteme” = knowledge) is the uncertainty about unique events resulting from insufficient knowledge, e.g. about whether a particular student will pass an exam or whether the increased release of CO_2 is the reason for the increasing global temperatures in the past 50 years. It does not make sense to imagine that a student takes the same exam many times, or that there are other earth systems where humans release CO_2 into the atmosphere. In such cases, the probability statements are based on an evaluation of the available facts and information by an individual and thus become subjective degrees of belief. Hence this approach is called the subjective or Bayesian approach (after Thomas Bayes, c.1701-1761). Subjectivity does not go well together with our ideal of scientific objectivity and thus this interpretation of probability is often criticized or rejected. However, subjectivity is not the same thing as arbitrariness – an individual should be able to reveal and explain the information on which his or her assignment of probabilites is based.

One can argue that aleatoric uncertainty is also a form of epistemic uncertainty about a physical system: If we knew exactly the position and the momentum of the dice when it leaves the hand of the player, then we should be able to predict the result. This is however not particularly helpful, and the distinction between repeatable and unique events remains.

One may ask whether we should measure the two types of uncertainty by the same concept. There is however an argument showing that in order to act rationally we should measure epistemic uncertainties by probabilities. I give this argument here. Assume that we have a finite set Ω of outcomes and that we are in the position of a bookmaker who has to fix payment odds for all subsets A of Ω other than Ω itself and the empty set. A payment odds $\lambda_A \in (0, \infty)$ for A is an offer for the following bet: A gambler who places b_A francs on A will win b_A/λ_A francs (in addition to the stake b_A that is returned) if A occurs and $-b_A$ francs if A does not occur. The stake b_A of the gambler can be negative: Because we do not make any assumptions on the payment odds, betting a negative amount on A

need not be equivalent to betting a positive amount on A^c. The net payoff of the gambler betting b_A on A is therefore

$$\phi_A(\omega) = 1_A(\omega)\frac{b_A}{\lambda_A} - (1 - 1_A(\omega))b_A = b_A\frac{1+\lambda_A}{\lambda_A}\left(1_A - \frac{\lambda_A}{1+\lambda_A}\right).$$

Clearly λ_A should reflect the bookmakers epistemic uncertainty about the occurence of A. The following Theorem due to de Finetti shows that if the bookmaker does not assign his payoff odds according to some probability π on Ω, then a gambler can combine bets on different events A in such a way that the total payoff is strictly positive no matter what the outcome ω is.

Theorem 1.1. *If there is a probability π on Ω such that $\lambda_A = \pi(A)/(1-\pi(A))$ for all A, then for any choice of bets (b_A), the payoff function $V(\omega) = \sum_A \phi_A(\omega)$ has expectation zero with respect to π and thus in particular takes both positive and negative values. If there is no such probability, then for any function $V : \Omega \to \mathbb{R}$ there is a choice of bets (b_A) such that $V(\omega) = \sum_A \phi_A(\omega)$ (in particular the gambler can place bets such that the payoff is equal to one for any outcome ω).*

Proof. The first statement follows by straightforward computation. For the second part, we use that the set of possible payoff functions is a linear subspace of $\mathbb{R}^\Omega$, namely the space spanned by all functions ϕ_A. If this subspace is not the whole space, then there is a nonzero vector π such that for all A

$$\sum_\omega \pi(\omega)\phi_A(\omega) = 0.$$

By the definition of ϕ_A, this means that for all A

$$\frac{1+\lambda_A}{\lambda_A}\sum_{\omega\in A}\pi(\omega) = \sum_{\omega\in\Omega}\pi(\omega).$$

By choosing as A all one-point sets, we see that $\sum_{\omega\in\Omega}\pi(\omega) \neq 0$ because otherwise π would be the zero vector. Without loss of generality, we assume that $\sum_{\omega\in\Omega}\pi(\omega) = 1$. Choosing again as A all one-point sets, we also obtain $\pi(\omega) > 0$ for all ω so that π is indeed a probability. Moreover,

$$\pi(A) = \sum_{\omega\in A}\pi(\omega) = \frac{\lambda_A}{1+\lambda_A}.$$

□

This theorem does not tell which probability π should be used to fix the payments odds. If we are in a situation with repeated outcomes according to some probability P, then in the long run a gambler can make a profit with high probability if $\pi \neq P$. But in such a situation the bookmaker would adjust the payment odds and thus also the underlying π based on past observations.

In the case where Ω is infinite and one has to show σ-additivity of π, there are some technical complications that I do not discuss here. See the unpublished note by D. A. Freedman, "Notes on the Dutch Book Argument" (2003), available at `http://www.stat.berkeley.edu/~census/` on which the theorem of this section is based.

1.2 Bayes formula

In the discrete case, we have a finite or countable partition (A_i) of Ω and another event $B \subset \Omega$ with $P(B) > 0$ (we tacitly assume that all sets we encounter are measurable). Then the following Bayes formula is well known and easy to prove

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{k:P(A_k)>0} P(B \mid A_k)P(A_k)} \quad (\text{if } P(A_i) > 0), \quad = 0 \quad (\text{if } P(A_i) = 0).$$

This formula has both a frequentist and a Bayesian interpretation. In the frequentist interpretation it gives the relative frequency of the event A_i among those repetitions where B occurs. In the subjective interpretation it tells how to modify the prior degree of belief $P(A_i)$ in A_i after observing that B has occured.

A somewhat counterintuitive implication of Bayes formula is the so-called base rate paradox: $P(B \mid A_i) \gg P(B \mid A_k)$ does not imply that $P(A_i \mid B) > P(A_k \mid B)$. It only holds that

$$\frac{P(A_i \mid B)}{P(A_k \mid B)} = \frac{P(B \mid A_i)}{P(B \mid A_k)} \frac{P(A_i)}{P(A_k)}$$

and the second factor on the right also influences the value on left. As an example, in case of a rare disease a positive outcome of a test with small error probabilities still can give a low probability of actually having the disease.

Various generalizations of Bayes' formula for continuous situations exist. If (X, Y) is a bivariate random vector with joint density $f_{X,Y}$, then we have the marginal densities

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

and the conditional densities

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

provided the denominator is neither zero nor infinity, and arbitary otherwise. Note that we are conditioning on an event with probability zero. Then it is obvious that

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|x')f_X(x')dx'}.$$

The densities above need not be with respect to Lebesgue measure, any product measure can be used. Moreover, the existence of densities for the marginal P_X of X is not needed: If the conditional distributions of Y given $X = x$ have densities $f_{Y|X}(y|x)$, then the conditional distributions of X given $Y = y$ are absolutely continuous with respect to the marginal P_X, and the density is given by

$$\frac{f_{Y|X}(y|x)}{\int f_{Y|X}(y|x')P_X(dx')}$$

for those y where the denominator is neither zero nor infinity and arbitrary otherwise.

1.3 Frequentist vs Bayesian statistics

1.3.1 Prior and posterior

Inferential statistics assumes that the observations x are realizations of a random vector X and aims to draw conclusions about the underlying distribution based on the observed values. The set of distributions $\mathcal{P}$ which are considered as possible are parametrized by θ: $\mathcal{P} = (P_\theta; \theta \in \Theta)$. In parametric statistics Θ is a subset of $\mathbb{R}^p$ and thus finite dimensional. We denote the space where the observations are by $\mathbf{X}$. Usually $\mathbf{X} = \mathbb{R}^n$ for some n.

Whereas frequentist statistics considers θ to be unknown, but fixed, Bayesian statistics treats θ as a random variable. Because repeated experiments with varying θ's drawn at random usually do not make sense, probability statements about θ have an epistemic meaning. The uncertainty about θ before seeing the data is described by a prior distribution π on Θ. The distribution P_θ is then the conditional distribution of X given θ, and we write the density of P_θ therefore as $f(x|\theta)$. The conditional density of θ given $X = x$ is called the posterior which we write as $\pi(.|x)$. The posterior describes therefore the uncertainty about θ after seeing the data. By Bayes formula the posterior density is

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{\int_\Theta \pi(\theta') f(x \mid \theta') d\theta'}.$$

By some abuse of notation we will denote marginal and conditional densities usually by f, without indicating with a subscript for which random variables. This will become clear from the arguments of f and the context. However, the prior density will always be denoted as $\pi(\theta)$ and the posterior as $\pi(\theta|x)$.

Example 1.1 (Normal means). *We discuss the case where we have n i.i.d. observations from a Gaussian distribution with unknown mean θ and known variance $\sigma^2 = 1$. If we choose $\mathcal{N}(\mu, \tau^2)$ as prior for θ, the density of the posterior is*

$$\begin{aligned}\pi(\theta|x_1, \dots, x_n) &= \frac{\prod_{i=1}^n f(x_i \mid \theta)\pi(\theta)}{\int \prod_{i=1}^n f(x_i \mid \theta')\pi(\theta') d\theta'} \\ &= \frac{\exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right)}{\int \exp\left(-\frac{1}{2\tau^2}(\theta' - \mu)^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \theta')^2\right) d\theta'}.\end{aligned}$$

By completing the square in the exponent, the numerator is seen to be equal to

$$\exp\left(-\frac{n + \tau^{-2}}{2}\left(\theta - \frac{\mu + n\tau^2\bar{x}}{1 + n\tau^2}\right)^2 - \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2 + \frac{n^2\tau^2}{2(1 + n\tau^2)}(\bar{x} - \mu)^2\right)$$

where $\bar{x}$ is the arithmetic mean $n^{-1}\sum x_i$. The same algebraic manipulation can be made in the integrand of the denominator. Because the second and third term in the exponent do not contain θ, they cancel and we are left with a Gaussian integral in the denominator. We therefore obtain

$$\pi(\theta|x_1, \dots, x_n) = \frac{\sqrt{n + \tau^{-2}}}{\sqrt{2\pi}} \exp\left(-\frac{n + \tau^{-2}}{2}\left(\theta - \frac{\mu + n\tau^2\bar{x}}{1 + n\tau^2}\right)^2\right),$$

or in other words

$$\theta \mid (X_1 = x_1, \dots, X_n = x_n) \sim \mathcal{N}\left(\frac{1}{1 + n\tau^2}\mu + \frac{n\tau^2}{1 + n\tau^2}\bar{x}, \frac{\tau^2}{1 + n\tau^2}\right).$$

The posterior mean is a convex combination of the prior mean μ and the maximum likelihood estimate $\bar{x}$ and the posterior variance is the prior variance divided by $1 + n\tau^2$.

The above computations could have been simplified by observing that in the numerator we can ignore any factor which does not contain θ because the same factor will appear in the denominator and thus cancels. Also the denominator is simply a normalization which ensures that the posterior density integrates to one. To emphasize this, one usually writes Bayes fomula as

$$\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta)$$

where the proportionality sign means up to a factor which does not contain θ, but may contain x. Because x is fixed and θ is variable, the second factor on the right is the likelihood and Bayes formula says that "the posterior is proportional to prior times likelihood".

The denominator $\int \pi(\theta) f(x \mid \theta) d\theta$ in Bayes formula for the posterior is the marginal density $f(x)$ of X, often also called the prior predictive density. The posterior predictive density for a future observation Y whose distribution depends on X and on θ is

$$f(y \mid x) = \int f(y \mid x, \theta) \pi(\theta | x) d\theta.$$

Example 1.2 (Normal means, ctd.). *The part in the numerator which does not depend on θ was found to be*

$$\exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n^2\tau^2}{1 + n\tau^2}(\bar{x} - \mu)^2\right)\right).$$

Using $\mathbf{1} = (1, \ldots, 1)^T$, the quadratic form in the exponent can be written as

$$(x - \mu\mathbf{1})^T \left(I - \frac{\tau^2}{1 + n\tau^2}\mathbf{1}\mathbf{1}^T\right)(x - \mu\mathbf{1}) = (x - \mu\mathbf{1})^T \left(I + \tau^2\mathbf{1}\mathbf{1}^T\right)^{-1}(x - \mu\mathbf{1}).$$

This shows that the prior predictive distribution of $(X_1, \ldots, X_n)$ is normal with mean $\mu\mathbf{1}$ and covariance matrix $\Sigma = I + \tau^2\mathbf{1}\mathbf{1}^T$. It can also be seen by writing $X_i = \theta + \varepsilon_i$ where $\theta \sim N(\mu, \tau^2)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ and all variables are independent. Finally, if we write a new observation X_{n+1} as $\theta + \varepsilon_{n+1}$, it follows that the posterior predictive density is

$$X_{n+1} \mid (X_1 = x_1, \ldots, X_n = x_n) \sim \mathcal{N}\left(\frac{1}{1 + n\tau^2}\mu + \frac{n\tau^2}{1 + n\tau^2}\bar{x}, 1 + \frac{\tau^2}{1 + n\tau^2}\right).$$

The choice of the prior is usually difficult and it can matter, see e.g. the base rate problem or the above example of a normal mean where the posterior can be any normal distribution if we choose the prior mean and variance accordingly. However, one is not allowed to let the prior depend on the data, it has to be chosen before seeing the data. The required choice of the prior is a weak point of Bayesian statistics which has lead to much controversy. We will discuss it in more detail in Chapter 2, including attempts to define non-informative priors.

1.3.2 Point estimation

As a Bayesian point estimate of θ, we can take any functional of location like expectation, median or mode of the posterior. If θ is a scalar, the posterior mean is

$$\widehat{\theta} = \mathbb{E}(\theta \mid x) = \frac{\int_{-\infty}^{\infty} \theta \pi(\theta) f(x \mid \theta) d\theta}{\int_{-\infty}^{\infty} \pi(\theta) f(x \mid \theta) d\theta},$$

the posterior median is the solution of

$$\int_{-\infty}^{\widehat{\theta}} \pi(\theta) f(x \mid \theta) d\theta = \frac{1}{2} \int_{-\infty}^{\infty} \pi(\theta) f(x \mid \theta) d\theta$$

and the posterior mode is

$$\widehat{\theta} = \arg\max_{\theta} p(\theta \mid x) = \arg\max_{\theta} \left(\log \pi(\theta) + \log f(x \mid \theta)\right).$$

(arg max denotes the argument where a function takes its maximal value). A unified approach which covers all these possibilities chooses a loss function $L : \Theta^2 \to [0, \infty)$ and minimizes the expected loss with respect to the posterior:

$$\widehat{\theta} = \arg\min_{t} \mathbb{E}(L(t, \theta) \mid x) = \arg\min_{t} \int_{-\infty}^{\infty} L(t, \theta) \pi(\theta \mid x) d\theta.$$

Here $L(t, \theta)$ quantifies our loss if the true value is θ and our estimate is t. We obtain the posterior mean if $L(t, \theta) = (t - \theta)^2$, the posterior median if $L(t, \theta) = |t - \theta|$ and the posterior mode if $L(t, \theta) = 1_{[-\varepsilon,\varepsilon]^c}(t - \theta)$ and we let ε go to zero.

In the frequentist approach, a point estimate $T(x)$ is evaluated in terms of the risk which is the expected loss with respect to the distribution of x

$$R(T, \theta) = \mathbb{E}_{\theta}(L(T(X), \theta)) = \int_{\mathbf{X}} L(T(x), \theta) f(x \mid \theta) dx$$

which depends on the unknown θ. We would like to minimize the risk for all θ simultaneously, but this is not possible because we can achieve a low risk for some value θ_0 if we are prepared to live with a high risk for values far away from θ_0 (consider the trivial estimator $T(x) \equiv \theta_0$, regardless of what the data are). To put it differently, if we have two estimators T_1 and T_2 which are both reasonable in an intuitive sense then we often have two values θ_1 and θ_2 such that

$$R(T_1, \theta_1) < R(T_2, \theta_1), \text{ but } R(T_1, \theta_2) > R(T_2, \theta_1).$$

In other words, whether T_1 or T_2 is better, depends on the unknown true value of the parameter θ.

In view of this dilemma, one solution is the minimax approach which looks for the estimator T which minimizes the maximal risk $\sup_{\theta} R(T, \theta)$.

A less pessimistic approach is to look for the estimator T which minimizes the the weighted risk

$$R(T, w) = \int_{\Theta} R(T, \theta) w(\theta) d\theta.$$

If $\int w(\theta) d\theta < \infty$, we can assume w.l.o.g. that w is a probability density. The next theorem shows that the estimator minimizing this weighted risk is then the Bayes estimator with prior $\pi = w$. Therefore we call $R(T, w)$ also the Bayes risk.

Theorem 1.2. *Assume that $\int w(\theta) d\theta = 1$ and choose w as the prior for θ. If*

$$T(x) = \arg\min_{t} \mathbb{E}(L(t, \theta) \mid x)$$

is well defined for almost all x with respect to the prior predictive distribution $f(x) = \int f(x \mid \theta) w(\theta) d\theta$, then T minimizes the weighted risk $R(T, w)$. Any other minimizer T' is almost surely equal to T.

Proof. By exchanging the order of integration, the weighted risk is

$$R(T, w) = \int_{\mathbf{X}} \left(\int_{\Theta} L(T(x), \theta) f(x|\theta) w(\theta) d\theta \right) dx$$

By Bayes formula $f(x|\theta)w(\theta) = f(x)\pi(\theta \mid x)$ and therefore

$$R(T, w) = \int_{\mathbb{X}} \left(\int_{\Theta} L(T(x), \theta) f(\theta \mid x) d\theta \right) f(x) dx.$$

In order to minimize the right-hand side, we have to minimize the inner integral for almost all x. □

In particular, we can avoid the discussion whether it is legitimate to consider θ as random using the prior only to weight the risk for the different values of θ. The posterior is then just a technical device to compute the estimator which minimizes the weighted risk.

A third approach to address the problem that there is no estimator which minimizes the risk simultaneously for all θ is admissibility. An estimator T is called admissible if no other estimator T' exists which is uniformly better than T: If $R(T', \theta) \leq R(T, \theta)$ for all θ, then we must have $R(T', \theta) = R(T, \theta)$ for all θ. This is allows to discard obviously bad estimators as inadmissible, but the class of admissible estimators is typically very large. It is not difficult to show that a Bayes estimator is admissible if the risk is continuous in θ for any estimator with finite risk and if the prior density is stricly positive everywhere. There is a large literature showing that under certain conditions any admissible estimator is a limit (in a sense to be made precise) of Bayes estimators. Such results give a justification of Bayes estimators from a frequentist point of view.

In the example of normal means the posterior mean $\widehat{\theta} = \frac{1}{1+n\tau^2}\mu + \frac{n\tau^2}{1+n\tau^2}\bar{x}$ is obviously biased:

$$\mathbb{E}_\theta(\widehat{\theta}) = \frac{1}{1+n\tau^2}\mu + \frac{n\tau^2}{1+n\tau^2}\theta \neq \theta \quad \forall \theta \neq \mu.$$

This is true in most cases because the choice of a prior usually means that not all values θ are considered equal. Since also modern frequentist statistics tends to deemphasize unbiasedness, this is however not a serious disadvantage of Bayesian estimators.

1.3.3 Testing and confidence intervals

Because in Bayesian statistics the parameter is random, it becomes possible to speak about the "probability that the null hypothesis is true" or the "probability that θ belongs to some interval". In frequentist statistics, such statements have no meaning, and one has to be very careful if one wants to explain the meaning of a p-value or a confidence interval in words.

Let us first discuss Bayesian testing of a null hypothesis $\theta \in \Theta_0 \subset \Theta$ against the alternative $\theta \in \Theta_1 = \Theta_0^c$. The posterior probability of the null hypothesis is then

$$\pi(\Theta_0 \mid x) = \int_{\Theta_0} \pi(\theta \mid x) d\theta = \frac{\int_{\Theta_0} f(x \mid \theta)\pi(\theta) d\theta}{\int_{\Theta} f(x \mid \theta)\pi(\theta) d\theta}.$$

A Bayesian test will reject the null hypothesis iff $\pi(\Theta_0 \mid x)$ is below some threshold c. If we quantify the loss in case of an error of the first kind as a_1 and in case of an error of the second kind as a_2, then the posterior expected loss of a test $\varphi : \mathbf{X} \to \{0, 1\}$

is $a_2(1 - \pi(\Theta_0 \mid x))$ if $\varphi(x) = 0$ and $a_1\pi(\Theta_0 \mid x)$ if $\varphi(x) = 1$. Hence the posterior expected loss is minimized if $\varphi(x) = 0$ for $\pi(\Theta_0 \mid x) > a_2/(a_1 + a_2)$ and $\varphi(x) = 1$ for $\pi(\Theta_0 \mid x) < a_2/(a_1 + a_2)$, i.e. $c = a_2/(a_1 + a_2)$.

Instead of $\pi(\Theta_0 \mid x)$, Bayesian statistics often considers the *Bayes factor* which is defined as the ratio of posterior and the prior odds in favor of the null hypothesis

$$B(x) = \frac{\pi(\Theta_0 \mid x)}{\pi(\Theta_1 \mid x)} \frac{\pi(\Theta_1)}{\pi(\Theta_0)}.$$

If $\Theta = \{\theta_0, \theta_1\}$ the Bayes factor is independent of the prior and equal to the likelihood ratio $f(x \mid \theta_0)/f(x \mid \theta_1)$ used in the Neyman-Pearson lemma. For composite hypotheses, the Bayes factor still depends on the prior. If

$$\pi_0(\theta) = \frac{\pi(\theta)1_{\Theta_0}(\theta)}{\pi(\Theta_0)}, \quad \pi_1(\theta) = \frac{\pi(\theta)1_{\Theta_1}(\theta)}{\pi(\Theta_1)}$$

denote the conditional priors under the null and the alternative, respectively, then

$$B(x) = \frac{\int_{\Theta_0} f(x \mid \theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x \mid \theta)\pi_1(\theta)d\theta} = \frac{f_0(x)}{f_1(x)} = \frac{f(x \mid \theta \in \Theta_0)}{f(x \mid \theta \in \Theta_1)},$$

the ratio of the marginal density of x under the null and the alternative hypothesis.

A Bayes factor between 1 and $\frac{1}{3}$ is considered as weak evidence against the null, a value between $\frac{1}{3}$ and 0.1 as substantial, a value between 0.1 and 0.01 as strong and a value below 0.01 as decisive, according to Jeffreys (1961).

In some applications the null hypothesis consists of a subset of Lebesgue measure zero, typically a lower dimensional subset of Θ, e.g. in the case of $\mathcal{N}(\mu, \sigma^2)$-observations $\Theta_0 = \{(\mu, \sigma^2); \mu = \mu_0\}$. In this case, if we choose a prior which has a density w.r. to Lebesgue measure, the prior and the posterior give zero probability to the null hypothesis. Hence there would be no need to collect data as data cannot change the prior belief in Θ_0. In such situations one should therefore choose a prior which assigns to Θ_0 a probability strictly between 0 and 1. This can be achieved by a mixture

$$\pi(d\theta) = \rho_0\pi_0(d\theta) + (1 - \rho_0)\pi_1(\theta)d\theta$$

where π_0 is a distribution which is concentrated on Θ_0 and ρ_0 is the prior probability of Θ_0. Because π_0 cannot have a density, we use here the general notation of measure theory. With such a prior, the posterior probability of Θ_0 is

$$\pi(\Theta_0 \mid x) = \frac{\rho_0 \int_{\Theta_0} f(x \mid \theta)\pi_0(d\theta)}{\rho_0 \int_{\Theta_0} f(x \mid \theta)\pi_0(d\theta) + (1 - \rho_0) \int_{\Theta} f(x \mid \theta)\pi_1(\theta)d\theta}.$$

In frequentist statistics, the p-value is taken as a measure of evidence for the null hypothesis. It is defined as the smallest significance level for which the null hypothesis is still rejected. Although conceptually this is not the same as the posterior probability of the null hypothesis, it would be nice if these two measures were close at least in the case where the null and the alternative are a priori equally likely. Let us consider the case where $\Theta_0 = \{\theta_0\}$. Then for $\rho_0 = \frac{1}{2}$

$$\pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \int_{\Theta} f(x \mid \theta)\pi_1(\theta)d\theta}$$

p-value	0.10	0.05	0.01	0.001
$\inf_{\pi_1} \pi(\Theta_0 \mid x)$	0.205	0.128	0.035	0.004
$\inf_{\pi_1} B(x)$	0.256	0.146	0.036	0.004
$\inf_{\pi_1 \in \mathcal{S}} \pi(\Theta_0 \mid x)$	0.392	0.290	0.109	0.018
$\inf_{\pi_1 \in \mathcal{S}} B(x)$	0.644	0.409	0.123	0.018

Table 1.1: Testing the null hypothesis $\mu = \mu_0$ for i.i.d normal observations with mean μ and known variance: Comparison of p-value and lower bounds on the posterior probability for the null under arbitrary (row 2) and under symmetric unimodal (row 3) priors for the alternative. The prior probability of the null hypothesis is equal to $\frac{1}{2}$. Source: Tables 4 and 6 in Berger and Selke, JASA 82 (1987)

which depends on the chosen prior for the alternative, but there is the trivial lower bound

$$\inf_{\pi_1} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_\theta f(x \mid \theta)}.$$

Numerical comparisons show that in many situations this lower bound is substantially larger than the p-value, see Table 1.1. This means that the p-value underestimates the evidence for the null (and therefore overestimates the evidence against the null) even when the prior is heavily biased towards the alternative. If we assume θ to be scalar and one restricts π_1 to the class $\mathcal{S}$ of symmetric unimodal densities, then one can show that

$$\inf_{\pi_1 \in \mathcal{S}} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_c \frac{1}{2c} \int_{\theta_0 - c}^{\theta_0 + c} f(x \mid \theta) d\theta}.$$

The discrepancy between the posterior probability of the null and the p-value is now even more drastic, see the third row of Table 1.1.

A Bayesian confidence set with level $1 - \alpha$ is called a $(1 - \alpha)$-*credible set.* It is a subset $C_x \subset \Theta$ (depending on x) such that

$$P(\theta \in C_x \mid X = x) = \pi(C_x \mid x) \geq 1 - \alpha.$$

Here x is fixed and θ is random, whereas in the frequentist approach θ is fixed and x is random and we require

$$P_\theta(C_x \ni \theta) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

(the different orders $\theta \in C_x$ and $C_x \ni \theta$ are chosen to emphasize what is random and what is fixed).

Among the many $(1-\alpha)$-credible sets, the one minimizing the volume (Lebesgue measure) is particularly attractive. It is obtained by taking C_x as a level set of the posterior, $C_x = L_{k_\alpha}$, where

$$L_k = \{\theta; \pi(\theta \mid x) \geq k\}, \quad k_\alpha = \sup\{k; \pi(L_k \mid x) \geq 1 - \alpha\}.$$

It is thus called a highest posterior density credible set. That it minimizes the volume can be seen as follows. For simplicity, we assume that $\pi(L_{k_\alpha} \mid x) = 1 - \alpha$. If C is another $(1 - \alpha)$-credible set, then by the definition of the level set

$$\begin{aligned} 0 \quad &\geq \quad \pi(L_{k_\alpha} \mid x) - \pi(C \mid x) = \int_{L_{k_\alpha} \cap C^c} \pi(\theta \mid x) d\theta - \int_{L_{k_\alpha}^c \cap C} \pi(\theta \mid x) d\theta \\ &\geq \quad k_\alpha \left(|L_{k_\alpha} \cap C^c| - |L_{k_\alpha}^c \cap C| \right) \end{aligned}$$

where $|C|$ denotes the volume (Lebesgue measure) of a set C. In high dimensions, the computation of L_{k_α} can be difficult.

1.3.4 Bayesian asymptotics

The two basic results in frequentist statistics are the following: Consider an i.i.d. model $X_i \sim f(x \mid \theta)dx$ where Θ is an open set in $\mathbb{R}^p$ and denote the likelihood function by $L_n(\theta) = \prod_{i=1}^n f(x_i \mid \theta)$, the maximum likelihood estimator (MLE) by $\widehat{\theta}_n = \arg\max_\theta L_n(\theta)$ and the Fisher information by

$$I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(X_i \mid \theta) \right).$$

Then, under regularity conditions, if θ_0 is the true parameter,

$$\widehat{\theta}_n \overset{\text{approx}}{\sim} \mathcal{N}(\theta_0, \frac{1}{n} I(\theta_0)^{-1})$$

and

$$2(\log L_n(\widehat{\theta}_n) - \log L_n(\theta_0)) \xrightarrow{d} \chi_p^2.$$

Bayesian asymptotics says – again under regularity conditions – that for any smooth prior which is strictly positive in a neighborhood of θ_0

$$\theta \mid (x_1, \ldots, x_n) \overset{\text{approx}}{\sim} \mathcal{N}(\widehat{\theta}_n, \frac{1}{n} I(\widehat{\theta}_n)^{-1}).$$

Therefore the influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{1/n}$ neighborhood of the MLE. There is a nice symmetry in the asymptotic statements, but note again the difference in what is considered fixed and what is random in the two approaches.

See for instance Schervish (1995), Chap. 7.4 for precise statements and proofs.

1.4 Likelihood Principle

We have seen in this chapter that a basic difference between the frequentist and the Bayesian approach is that the former considers other values of the data that did not occur, but could have been obtained, whereas the latter considers only the data that were actually observed. To justify the frequentist approach, we mention the following quote from Mosteller and Tukey (1968): “One hallmark of the statistically conscious investigator is his firm belief that however the survey, experiment or observational program actually turned out, it could have turned out somewhat differently. Holding such a belief and taking appropriate actions make effective use of data possible. We need not always ask explicitly “How much differently ?”, but we should be aware of such questions.”

On the other hand, there are clearly situations where considering data that were not, but might have been obtained leads to strange conclusions: Consider estimation of an unknown concentration of a substance in a probe. The probe can be analyzed by two labs, one which measures with high precision and one with low precision. The high precision lab is however not always available, due to high demand from other customers. Let us assume that the standard deviations of both labs are known and equal to 1 and 10 respectively, and that the chance the precise lab is available is 0.5. If the analysis was made by the imprecise lab, then I only know that the true value is within ±19.6 from the result. Arguing that because there was a 50% chance to have the analysis done in the precise lab and that therefore the standard deviation is $\sqrt{0.5 \cdot 1^2 + 0.5 \cdot 10^2} = 7.1$ is obviously not reasonable.

This can be formulated as the
Conditionality principle: If an experiment for inference about a parameter θ is chosen independently from a collection of different possible experiments, then any experiment not chosen is irrelevant to the inference.

This principle seems quite indisputable. There is another principle which is also not controversial, namely that observations which differ only in a way which is irrelevant for the model under consideration should lead to the same conclusion. For instance, in a model where the observations X_i are i.i.d. $\sim f(x|\theta)dx$, the time order of the observations is irrelevant. The time order can give information whether the i.i.d. assumption is reasonable or not, but once I decide to use the model, the time order of the observation does not give any information about θ. In mathematical statistics this idea is formalized by the concept of a *sufficient statistic*: A function $T(x)$ of the observation x is sufficient for a model $(f(x|\theta); \theta \in \Theta)$ if the conditional distribution of x given $T(x) = t$ does not involve θ. Then we can formulate

Sufficiency principle: If there are two observations x and y such that $T(x) = T(y)$ for a sufficient statistic T, then any conclusion about θ should be the same for x and y.

The surprising result due to Birnbaum (1962) is that the two largely non-controversial principles above imply a third principle which is violated by many frequentist tests and confidence intervals:

Likelihood principle: If there are two different experiments for inference about the same parameter θ and if the outcomes x and y from the two experiments are such that the likelihood functions differ only by a multiplicative constant, then the inference should be the same.

To understand this principle, let us consider the problem where there is some event with unknown probability p and I want to test the null hypothesis $p \leq 0.5$ against the alternative $p > 0.5$. This can be done either by repeating the trial n times and observing the number X of trials where the event occured, or by repeating the experiment a random number N times until the event has occured a fixed number of times x. In the first experiment X is random, n is fixed,

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

and and the null hypothesis is rejected if $x/n > c_1$. In the second experiment N is random, x is fixed and

$$P_p(N = n) = \binom{n-1}{x-1} p^x (1-p)^{n-x}$$

because the last occurence of the event must be in the n-th trial, but the other $x-1$ occurences can be any time before the n-th trial. In the second experiment, we will reject the null if $n < c_2$ which is equivalent to $x/n > x/c_2 =: c_3$ (remember that x is fixed). Hence the form of the rejection region is the same in both experiments, but the boundary values c_1 and c_3 differ in general because they are computed differently:

$$\sum_{k=c_1}^{n} \binom{n}{k} 2^{-n} > \alpha \geq \sum_{k=c_1+1}^{n} \binom{n}{k} 2^{-n},$$

but

$$\sum_{m=x}^{c_2-1} \binom{m-1}{x-1} 2^{-m} \leq \alpha < \sum_{m=x}^{c_2} \binom{m-1}{x-1} 2^{-m}.$$

Hence for a frequentist test it matters whether I observe x successes in n trials in experiment 1 or experiment 2 although the likelihood functions $p \mapsto P_p(X = x)$ and $p \mapsto P_p(N = n)$ are in both cases proportional to $p^x(1-p)^{n-x}$. In particular the Bayes test gives the same answer regardless which experiment was performed.

Note that the maximum likelihood estimator is x/n, regardless whether experiment 1 or 2 was made: Point estimation by maximum likelihood does obey the likelihood principle.

For a proof of Birnbaum's result, see for instance Section 1.3.3 in Robert's book.

Chapter 2

Prior distributions

The choice of a prior is a point which has lead to an intensive debate and which is often considered to be the weak point of the Bayesian approach. We discuss three approaches: The first one chooses prior distributions such that the posterior can be easily computed. The second one tries to determine a prior which contains as little information as possible. The third one tries to choose a prior based on the opinion of one or several experts.

2.1 Conjugate priors

In the example 1.1 a normal prior lead to a normal posterior and therefore the application of Bayes formula becomes particularly simple: We need to know only how to compute the mean and variance of the posterior. The following definition generalizes this feature.

Definition 2.1. *A parametric family $\mathcal{P}_\Xi = \{\pi_\xi(\theta); \xi \in \Xi\}$, $\Xi \subset \mathbb{R}^q$, of prior densities is called* conjugate *for the model $\{f(x \mid \theta); \theta \in \Theta\}$ if, for any $\pi \in \mathcal{P}_\Xi$ and any x, $\pi(\theta \mid x)$ is again in $\mathcal{P}_\Xi$.*

Written out, this means that to any $\xi \in \Xi$ and any x there must be a $\xi' = \xi'(\xi, x)$ such that

$$\pi_\xi(\theta) f(x \mid \theta) \propto \pi_{\xi'}(\theta).$$

Computing the posterior amounts then to computing $\xi'(\xi, x)$.

It is obvious that $\mathcal{P}_\Xi$ is conjugate if the following two conditions are satisfied

1. To any x there is a $\xi(x) \in \Xi$ such that $f(x \mid \theta) \propto \pi_{\xi(x)}(\theta)$.

2. To any pair $\xi_1, \xi_2 \in \Xi$ there is a $\xi_3 \in \Xi$ such that $\pi_{\xi_1}(\theta)\pi_{\xi_2}(\theta) \propto \pi_{\xi_3}(\theta)$.

Example 2.1. *Consider the binomial dstribution*

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

It is clear that the above conditions are satisfied if we choose $\pi_\xi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ where $\xi = (\alpha, \beta) \in \mathbb{N}^2$ or $\xi \in (0, \infty)^2$.

A class of conjugate priors $\mathcal{P}_\Xi$ remains conjugate under repeated sampling, i.e. it is also conjugate for the model where $X_1, \ldots, X_n$ are i.i.d., $X_i \sim f(x \mid \theta)dx$ and n is arbitrary because for instance

$$\pi(\theta \mid x_1, x_2) \propto \pi(\theta \mid x_1) f(x_2 \mid \theta).$$

Therefore, if $\mathcal{P}_\Xi$ is conjugate for $f(x \mid \theta)$, for arbitrary, but fixed ξ_0 we can write

$$\prod_{i=1}^n f(x_i \mid \theta) = \frac{\pi_{\xi_n(x_1,\ldots,x_n)}(\theta)}{\pi_{\xi_0}(\theta)} f_n(x_1, \ldots, x_n)$$

where f_n is the prior predictive density of $X_1, \ldots, X_n$ and where ξ_n maps n-tupels of observed values to Ξ. In the language of mathematical statistics, ξ_n is then a sufficient statistic whose dimension is the same for any n. The existence of a sufficient statistics of finite dimension has been studied in mathematical statistics, and unfortunately it exists only for a restricted class of models, thus limiting the use of conjugate priors to such models. If the set $\{x; f(x \mid \theta) > 0\}$ does not depend on θ, f must belong to a so-called exponential family where the densities have the following form

$$f(x \mid \theta) = \exp(c_1(\theta)T_1(x) + \ldots c_q(\theta)T_q(x) + d(\theta))h(x).$$

The conjugate family consists then of densities

$$\pi_\xi(\theta) \propto \exp(c_1(\theta)\xi_1 + \ldots c_q(\theta)\xi_q + d(\theta)\xi_{q+1}).$$

The table below gives examples of exponential families together with their conjugate prior distributions.

Model	prior	posterior
Binomial(n, θ)	Beta(α, β)	Beta$(\alpha + x, \beta + n - x)$
Multinomial $(n, \theta_1, \ldots, \theta_k)$	Dirichlet$(\alpha_1, \ldots, \alpha_k)$	Dirichlet$(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
i.i.d. Poisson(θ)	Gamma(γ, λ)	Gamma$(\gamma + \sum_i x_i, \lambda + n)$
i.i.d. Normal$(\mu, \frac{1}{\tau})$ $\theta = (\mu, \tau)$	Normal$(\mu_0, \frac{1}{n_0 \tau}) \times$ Gamma(γ, λ)	Normal$(\frac{n}{n+n_0}\bar{x} + \frac{n_0}{n+n_0}\mu_0, \frac{1}{(n+n_0)\tau}) \times$ Gamma$(\gamma + \frac{n}{2}, \lambda + \frac{1}{2}\sum_i (x_i - \bar{x})^2 + \frac{n n_0}{2(n+n_0)}(\bar{x} - \mu_0)^2)$
Uniform$(0, \theta)$	Pareto(α, σ)	Pareto$(\alpha + n, \max(\sigma, x_1, \ldots, x_n))$

Table 2.1: Most frequently used models with their conjugate priors

The distributions which appear in this table are defined as follows

- Beta(α, β) where $\alpha > 0$ and $\beta > 0$ is the distribution on $(0, 1)$ with the density

  $$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}.$$

  Here $\Gamma(\alpha) = \int_0^\infty x^{\alpha - 1} e^{-x} dx$ is the Gamma function.

- Gamma(γ, λ) where $\gamma > 0$ and $\lambda > 0$ is the distribution on $(0, \infty)$ with the following density

  $$f(x) = \frac{\lambda^\gamma}{\Gamma(\gamma)} x^{\gamma - 1} e^{-\lambda x}.$$

- Multinomial$(n, p_1, \ldots, p_k)$ where $p_i \geq 0$, $\sum_i p_i = 1$ is the discrete distribution on the set $\{(x_1, \ldots, x_k) : x_i \geq 0, \sum_{i=1}^k x_i = n\}$ given by

$$p(x_1, \ldots, x_k) = \frac{n!}{(x_1)! \ldots (x_k)!} p_1^{x_1} \ldots p_k^{x_k}.$$

- Dirichlet$(\alpha_1, \ldots, \alpha_k)$ is the distribution on the simplex $\{(p_1, \ldots, p_{k-1}); p_i \geq 0, \sum_i p_i \leq 1\}$ with density

$$f(p_1, \ldots, p_{k-1}) = \frac{\Gamma(\alpha_1 + \ldots, \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} \cdots p_{k-1}^{\alpha_{k-1}-1} (1 - p_1 - \ldots p_{k-1})^{\alpha_k - 1}.$$

  For $k = 2$ this is simply the Beta-distribution.

- Pareto(α, σ) $(\alpha > 0, \sigma > 0)$ is the distribution on $[\sigma, \infty)$ with density

$$f(x) = \alpha \sigma^\alpha x^{-(\alpha+1)}.$$

In the case of the normal distribution, the bivariate prior for the mean μ and the inverse of the variance, the so-called precision τ is specified as conditional distribution of μ given τ (the first line in the table) times the marginal of τ (the second line in the table). Written out, the joint prior density is

$$\pi(\mu, \tau) = \frac{\sqrt{n_0 \tau}}{\sqrt{2\pi}} \exp\left(-\frac{n_0 \tau}{2}(\mu - \mu_0)^2\right) \frac{\lambda^\gamma}{\Gamma(\gamma)} \tau^{\gamma - 1} \exp(-\lambda \tau).$$

The formula for the posterior follows therefore by multiplying the prior with the likelihood and using the same kind of algebraic manipulations as in Example 1.1. Integrating over τ we obtain as marginal prior for μ the density

$$\pi(\mu) \propto \int_0^\infty \tau^{\gamma - 0.5} \exp\left(-\tau(\lambda + \frac{n_0}{2}(\mu - \mu_0)^2)\right) d\tau \propto \left(1 + \frac{n_0}{2\lambda}(\mu - \mu_0)^2\right)^{\gamma + 0.5}$$

i.e. the marginal prior for μ is a shifted and scaled t-distribution with $\gamma + 1/2$ degrees of freedom. Similarly, the marginal posterior of μ is a shifted and scaled t-distribution with $\gamma + (n+1)/2$ degrees of freedom.

There are also analogous conjugate distributions for the d-dimensional normal distribution $\mathcal{N}_d(\mu, \Psi^{-1})$ with unknown mean vector μ and unknown precision matrix Ψ. The marginal of Ψ is then the so-called Wishart distribution, and the conditional distribution of μ given Ψ is multivariate normal. We refer to the literature for the definitions and the formulae.

Conjugate priors have again parameters, usually called *hyperparameters*, which have to be chosen (In the general formula the hyperparameters are called ξ_i, in Table 2.1 different symbols are used). Hence using a conjugate prior does not answer the question "which prior ?". As typically there are more hyperparameters than parameters, choosing a hyperparameter seems even more difficult than choosing a parameter value. Note however that usually one of the hyperparameters can be regarded as a hypothetical sample size of the prior: In the general case, it is the parameter ξ_{q+1}, in the binomial case it is $\alpha + \beta$, in the Poisson case it is γ, in the normal case it is n_0 for μ and γ for τ. So this parameter can be determined by asking how much we want to rely on the prior. The other parameters usually are related to a location parameter of the prior which helps to choose its value. For instance in the case of a Beta distribution, the mean is $\alpha/(\alpha + \beta)$.

2.2 Noninformative priors

The search for a noninformative prior is motivated by the wish to reduce the subjective element in a Bayesian analysis. A first attempt defines a uniform prior on Θ to be noninformative: This has however two drawbacks. First, the uniform distribution on Θ is a probability only if Θ has finite volume. Second, the uniform distribution is not invariant under different ways to parametrize the same family of distribution. Assume we use $\tau = g(\theta)$ as our new parameter where g is invertible and smooth, i.e. $\theta = g^{-1}(\tau)$. If θ has density π, then by a standard result from multivariate calculus, $\tau = g(\theta)$ has the density

$$\lambda(\tau) = \pi(g^{-1}(\tau))|\det Dg(g^{-1}(\tau))|^{-1},$$

where Dg is the matrix whose (ij)-th entry is $\partial g_i/\partial\theta_j$, the so-called Jacobi matrix. Hence if π is constant and g is not linear, then λ is not constant.

As we shall see soon, also other approaches often lead to priors which are of the form $\pi(\theta) \propto g(\theta)$ where $\int_\Theta g(\theta)d\theta = \infty$. Such priors are called improper, and we will discuss them briefly in a separate subsection below.

2.2.1 Jeffreys prior

Jeffreys proposed to take

$$\pi(\theta) \propto \det(I(\theta))^{1/2}$$

where $I(\theta)$ is the Fisher information matrix that we already encountered in 1.3.4. It is defined as

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(X_i \mid \theta)\right),$$

and one can show that it is also equal to

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i \mid \theta)\left(\frac{\partial}{\partial\theta}\log f(X_i \mid \theta)\right)^T\right).$$

The rationale for this proposal is the following: Because $I(\theta)^{-1}$ is the asymptotic variance of the MLE, $\det(I(\theta))^{-1/2}$ is proportional to the volume of a set of parameters that cannot be distinguished based on the data. Hence Jeffreys prior gives approximately equal probabilities to regions with an approximately equal number of "distinguishable parameters".

Example 2.2. *Normal distribution. If $X \sim \mathcal{N}(\theta, 1)$, then $I(\theta) = 1$, so Jeffreys prior is the improper uniform distribution on $\mathbb{R}$. If $X \sim \mathcal{N}(0, \theta^2)$, $I(\theta) = 2/\theta^2$, so Jeffreys prior is $\pi(\theta) \propto \theta^{-1}$, which is again improper. It is the uniform distribution for the parameter $\tau = \log\theta$. If both mean and standard deviation are unknown, $X \sim (\mu, \sigma)$ and $\theta = (\mu, \sigma)$, then the Fisher information is diagonal with elements σ^{-2} and $2\sigma^{-2}$ and therefore Jeffreys prior is proportional to σ^{-2}. In particular, it is not the product of the univariate Jeffreys priors for μ and σ, contrary to what one would expect.*

Example 2.3. *Binomial distribution. If X is binomial(n, θ), then*

$$I(\theta) = \mathbb{E}_\theta\left(\frac{X}{\theta^2} + \frac{n-X}{(1-\theta)^2}\right) = \frac{n}{\theta(1-\theta)}$$

so Jeffreys prior is the Beta$(\frac{1}{2}, \frac{1}{2})$ and not the uniform distribution. In particular it is proper.

We now show that Jeffreys prior is equivariant under different parametrizations. We consider first the example $X \sim \mathcal{N}(0, \sigma^2)$ and the four parametrizations $\theta = \sigma$ (standard deviation), $\theta = \sigma^2$ (variance), $\theta = \sigma^{-2}$ (precision) amd $\theta = \log(\sigma)$. The results for these four choices are summarized in Table 2.2

Parameter	$\frac{\partial^2 \log f(x\|\theta)}{\partial\theta^2}$	$I(\theta)$	Jeffreys prior	$\frac{d\sigma}{d\theta}$
$\theta = \sigma$	$\frac{\theta^2 - 3x^2}{\theta^4}$	$\frac{2}{\theta^2}$	$\propto \frac{1}{\theta} = \frac{1}{\sigma}$	1
$\theta = \sigma^2$	$\frac{\theta - 2x^2}{2\theta^3}$	$\frac{1}{2\theta^2}$	$\propto \frac{1}{\theta} = \frac{1}{\sigma^2}$	$\frac{1}{2\sqrt{\theta}}$
$\theta = \sigma^{-2}$	$-\frac{1}{2\theta^2}$	$\frac{1}{2\theta^2}$	$\propto \frac{1}{\theta} = \sigma^2$	$-\frac{1}{2\theta^{3/2}}$
$\theta = \log(\sigma)$	$-2e^{2\theta}x^2$	2	$\propto 1$	e^{θ}

Table 2.2: Fisher information and Jeffreys prior for different parametrizations of $\mathcal{N}(0, \sigma^2)$

If σ has the density $\frac{1}{\sigma}$, then by the rules of calculus, $\theta = \theta(\sigma)$ has the density $\frac{1}{\sigma(\theta)}|\frac{d\sigma}{d\theta}|$. So for instance $\theta = \sigma^{-2}$ has the density $\sqrt{\theta}\frac{1}{2\theta^{3/2}} = \frac{1}{2\theta} = \frac{\sigma^2}{2}$ which shows the equivariance. The other cases can be checked similarly.

The argument in the general multiparameter case $\tau = g(\theta)$ goes as follows. Denoting the Jacobi matrices of g and g^{-1} by Dg and Dg^{-1}, respectively, the chain rule implies

$$\frac{\partial}{\partial\tau} \log f(x \mid g^{-1}(\tau)) = (Dg^{-1}(\tau))^T \frac{\partial}{\partial\theta} \log f(x \mid g^{-1}(\tau)).$$

Hence by the second form of the Fisher information, the Fisher information with respect to τ is

$$I_\tau(\tau) = (Dg^{-1}(\tau))^T I_\theta(g^{-1}(\tau)) Dg^{-1}(\tau).$$

Moreover, $Dg^{-1}(\tau)$ is the inverse matrix of $Dg(g^{-1}(\tau))$ and therefore Jeffreys prior for τ is proportional to

$$\det(I_\tau(\tau))^{1/2} = |\det Dg(g^{-1}(\tau))|^{-1} \det(I_\theta(g^{-1}(\tau)))^{1/2},$$

in accordance with the above result for the transformation of densities.

For a scalar parameter, Jeffreys prior is usually a good choice, although it violates the likelihood principle because the Fisher information contains an integral over X. However, for vector parameters, it can have undesirable features.

Example 2.4. *Vector normal means: Consider the model where X_i $1 \le i \le 2n$ are independent and normally distributed with variance σ^2 and means $\mathbb{E}(X_{2k-1}) = \mathbb{E}(X_{2k}) = \mu_k$. Jeffreys prior for this model is $\pi(\mu_1, \ldots, \mu_n, \sigma) \propto \sigma^{-n-1}$, and the posterior is proportional*

$$\sigma^{-2n-1} \exp\left(-\frac{1}{4\sigma^2}\sum_{k=1}^{n}(x_{2k} - x_{2k-1})^2\right) \cdot \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(\mu_k - \frac{x_{2k} + x_{2k-1}}{2}\right)^2\right)$$

The second factor integrated with respect to $\mu_1, \ldots, \mu_n$ gives $(2\pi)^{n/2}$. Therefore the first factor is proportional to the marginal posterior of σ, and one finds that

$$\mathbb{E}(\sigma^2 \mid x_1, \ldots, x_{2n}) = \frac{1}{4(n-1)}\sum_{k=1}^{n}(x_{2k} - x_{2k-1})^2.$$

As $n \to \infty$, this converges to $\frac{\sigma^2}{2}$. In other words, Jeffreys prior leads to a Bayes estimator with unsatisfactory frequentist properties.

2.2.2 Improper priors

If π has infinite total mass, $\pi(\theta)f(x \mid \theta)$ can have finite or infinite total mass, depending on the likelihood. If the total mass is finite, then we have by formal analogy the posterior density

$$\pi(\theta \mid x) = \frac{\pi(\theta)f(x \mid \theta)}{\int \pi(\theta')f(x \mid \theta')d\theta'}$$

and we can construct Bayesian point estimates, tests and credible intervals as before. Typically, this can be justified by approximating the improper prior by a sequence of proper priors π_k and showing that the associated sequence of posteriors $\pi_k(\theta \mid x)$ converges to the above expression. However, even if this convergence holds, paradoxes can occur. Moreover, in complicated models it is not always easy to check whether $\pi(\theta)f(x \mid \theta)$ has finite total mass.

2.2.3 Reference priors

The concept of reference priors was introduced by Bernardo (J. Roy. Statist. Soc. 41, 1979) and it has been studied by Berger and Bernardo in a series of papers afterwards. It has two aspects: A new justification of Jeffeys prior and a distinction between parameters of interest and nuisance parameters.

We begin with the former and call a prior π noninformative if the difference between the prior π and the posterior $\pi(. \mid x)$ in some distance is maximized. This seems reasonable because if the data x have the largest possible impact, the impact of the prior is minimal. In the opposite case, the prior is most informative if it is a point mass at some value θ_0 and then the prior and the posterior coincide. There are however two problems with this idea, the choice of the distance and the dependence on the data x. Bernardo proposed to use Kullback-Leibler divergence as the distance measure and to integrate over the data according to the prior predictive distribution $f(x) = \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$. The Kullback-Leibler divergence between to densities f and g is defined as

$$K(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx.$$

Because in general $K(f, g) \neq K(g, f)$, it is not a true distance, but it satisfies $K(f, g) \geq 0$ and $K(f, g) = 0$ iff $f(x) = g(x)$ for almost all x. Bernardo calls π a reference prior if it maximizes

$$\begin{aligned} I(X, \Theta) &= \int_X f(x) \int_\Theta \pi(\theta \mid x) \log \frac{\pi(\theta \mid x)}{\pi(\theta)} d\theta dx = \int_\Theta \pi(\theta) \int_X f(x \mid \theta) \log \frac{\pi(\theta)f(x \mid \theta)}{\pi(\theta)f(x)} dx d\theta \\ &= \int_\Theta \pi(\theta) \int_X f(x \mid \theta) \log \pi(\theta \mid x) dx d\theta - \int_\Theta \pi(\theta) \log \pi(\theta) d\theta. \end{aligned}$$

In information theory $I(X, \Theta)$ is called the mutual information of X and θ, confirming the idea behind this approach: A prior is most noninformative if the mutual information of X and θ is maximal. Finding the maximizer is however complicated and there is no closed form solution because f also depends on π. The problem becomes much simpler if we look at n i.i.d. observations with density $f(x \mid \theta)$ and let n go to infinity because then the posterior π becomes independent of the prior: As stated in Subsection 1.3.4, it holds

$$2 \log \pi(\theta \mid x_1, \ldots, x_n) \approx -p \log(2\pi) + \log \det(nI(\widehat{\theta})) - n(\widehat{\theta} - \theta)^T I(\widehat{\theta})(\widehat{\theta} - \theta)$$

where $\widehat{\theta}$ is the MLE and p is the dimension of θ. By the result on the (frequentist) asymptotics of the MLE from the same subsection, $\widehat{\theta}$ is approximately $\mathcal{N}(\theta, \frac{1}{n}I(\theta)^{-1})$. Hence

$$\int \log\det(nI(\widehat{\theta}))\prod_{i=1}^{n} f(x_i \mid \theta)dx_i \approx p\log n + I(\widehat{\theta}))$$

and

$$n\int (\widehat{\theta}-\theta)^T I(\widehat{\theta})(\widehat{\theta}-\theta)\prod_{i=1}^{n} f(x_i \mid \theta)dx_i \approx p.$$

Therefore we obtain that

$$\int \log\pi(\theta \mid x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i \mid \theta)dx_i \approx \frac{p(\log n - \log(2\pi) - 1) + \log\det I(\theta)}{2}.$$

This finally gives the approximation

$$I((X_1,\ldots,X_n),\Theta) \approx \frac{p(\log n - \log(2\pi) - 1)}{2} + \int_{\Theta} \pi(\theta)\log\frac{\det I(\theta)^{1/2}}{\pi(\theta)}d\theta.$$

The integral on the right hand side is maximal for $\pi(\theta) = c^{-1}\det I(\theta)^{1/2}$ because by concavity of the log it follows that $\log z \leq \log c + (z-c)/c$ with equality iff $z = c$. We thus have obtained again Jeffreys prior in the limit.

Next, we discuss briefly Bernardo's approach if $\theta = (\theta_1, \theta_2)$ can be decomposed in the parameter of interest θ_1 is and the nuisance parameter θ_2. In this case, the prior is factorized as $\pi(\theta) = \pi(\theta_2 \mid \theta_1)\pi(\theta_1)$ and one $\pi(\theta_2 \mid \theta_1)$ as the Jeffreys prior for the model with fixed value θ_1 and $\pi(\theta_1)$ as the Jeffreys prior for the model

$$f^*(x \mid \theta_1) = \int_{\Theta_2} f(x \mid \theta)\pi(\theta_2 \mid \theta_1)d\theta_2.$$

If $\pi(\theta_2 \mid \theta_1)$ is not proper, $f^*(x \mid \theta)$ is not a probability density and one needs to restrict θ_2 first to a compact subset. In the end, one then considers the limit of the prior $p(\theta_1)$ as this compact subset tends to the whole space.

In the example 2.4 with normal vector means, we saw that Jeffreys prior leads to an undesirable estimate of σ^2. To correct this, we consider σ as the parameter of interest θ_1 and $\mu_1,\ldots,\mu_n$ as the nuisance parameter θ_2. Then $\pi(\mu_1,\ldots,\mu_n \mid \sigma)$ is flat, and

$$f^*(x_1,\ldots,x_{2n} \mid \sigma) \propto \sigma^{-n}\exp\left(-\frac{1}{4\sigma^2}\sum_{k=1}^{n}(x_{2k}-x_{2k-1})^2\right).$$

Although this "density" has infinite total mass, a formal calculation gives $\pi(\sigma) \propto 1/\sigma$. Finally, one can check that the prior $\pi(\mu_1,\ldots,\mu_n,\sigma) \propto 1/\sigma$ gives

$$\mathbb{E}(\sigma^2 \mid x_1,\ldots,x_{2n}) = \frac{1}{2(n-2)}\sum_{k=1}^{n}(x_{2k}-x_{2k-1})^2$$

which is much more reasonable than what we had with Jeffreys prior.

2.3 Expert priors

If one takes the view seriously that the prior quantifies knowledge and uncertainty prior to seeing the data, then it makes sense to try to elicit a prior from one or several experts. How to do this best has become a research topic in itself which is situated at the intersection of statistics and psychology. One of the conclusions of research in this area is that expert judgement is subject to various kinds of heuristics and biases, and that the size of unwanted effects depends strongly on how questions are phrased. A standard procedure in the case of a univariate prior is to first elicit a number of summary statistics. In a second step one then fits a distribution which takes these summaries into account. Taking as summary statistics the median and the quartiles or the 33% and 67% quantiles seems to be a good choice: Expectation, standard deviation and extreme quantiles are more difficult to elicit. Eliciting dependence in multivariate priors is considerably more difficult.

For a more detailed discussion of this issue, see e.g. Garthwaite, Kadane and O'Hagan, J. Amer. Statist. Assoc. 100 (2005), 680-700.

2.4 Discussion

The justification of a chosen prior remains a difficult issue. The concept of a noninformative prior is difficult to implement in complex models with many parameters. Therefore, some subjective choices are often unavoidable. However, in most cases any reasonable choice leads to similar conclusions because the likelihood tends to dominate – at least if the number of observations is large (compare section 1.3.4). In any practical application, one should check that at least marginally the prior is approximately constant in a highest probability density credible set, or do a sensitivity analysis by varying the prior. Choosing a prior such that the desired conclusion is obtained is always possible, but this would be unethical and such a choice would be hard to justify.

The choice of the prior for linear regression models in combination with variable selection – an important practical case – will be discussed in Section 3.2.

If the number of parameters is large compared to the number of observations, then the prior often matters. This seems however unavoidable. In that situation, frequentist statistics often uses regularization methods which usually have a Bayesian interpretation. For instance if we use penalized maximum likelihood estimation

$$\widehat{\theta} = \arg\max(\log f(x \mid \theta) + P(\theta))$$

the penalty $P(\theta)$ can usually be interpreted as the log of a prior density.

Chapter 3

Hierarchical Bayes models

3.1 Hierarchical models

Until now we distinguished only between parameters θ and observations x. They are both random with a joint distribution that is specified by the marginal for θ and the conditional distribution of x given θ. In hierarchical models the prior for θ depends on other parameters ξ, called hyperparameters which are given again a prior distribution. This leads to a triple of random variables (ξ, θ, x) with joint density

$$\pi(\xi)\pi(\theta \mid \xi)f(x \mid \theta).$$

The basic approach of Bayesian statistics remains unchanged: Once we have specified the three factors of the joint distribution of the triple (ξ, θ, x), we compute the posterior, that is the conditional distribution of the unobserved variables (ξ, θ) given the observed variables by the rules of probability, and then we base our conclusions on this posterior. Often, the primary interest is in the original parameter θ and then we need the marginal posterior $\pi(\theta \mid x)$. There are two ways to compute it.

In the first approach, we begin by computing the marginal prior

$$\pi(\theta) = \int \pi(\theta \mid \xi)\pi(\xi)d\xi$$

and then use Bayes formula

$$\pi(\theta \mid x) \propto \pi(\theta)f(x \mid \theta).$$

This shows in particular that the introduction of hyperparameters is equivalent to a special choice of a prior for θ.

There are however situations where the approach based on the following formula is computationally easier:

$$\pi(\theta \mid x) = \int \pi(\theta \mid x, \xi)\pi(\xi \mid x)d\xi \propto \int \pi(\theta \mid x, \xi)\pi(\xi)f(x \mid \xi)d\xi$$

(the first equality is the law of total probability applied to the conditional distribution given x). In conjugate situations we have an explicit expression not only for $\pi(\theta \mid x, \xi)$, but usually also for $f(x \mid \xi)$ as we shall show in the examples below. The final integration over ξ usually cannot be done in closed form, but will need some approximation. Such approximations will be the topic of Chapter 4.

Example 3.1. *Normal means. Let $X_1, \dots X_n$ be i.i.d. and $\mathcal{N}(\theta, 1)$-distributed, and consider the prior $\theta \sim \mathcal{N}(\mu, \tau^2)$. The posterior is then*

$$\theta \mid x_1, \dots, x_n \sim \mathcal{N}\left(\frac{1}{1+n\tau^2}\mu + \frac{n\tau^2}{1+n\tau^2}\bar{x}, \frac{\tau^2}{1+n\tau^2}\right),$$

see Example 1.1. This is sensitive to the choice of τ^2, so we can consider τ^2 as hyperparameter to which we assign a distribution. (Of course, the posterior is also sensitive to the choice of μ, but if we want to be noninformative with respect to the prior location, we would take Jeffreys prior $\pi(\theta) \propto 1$). The marginal prior is then a scale mixture of normal priors which is often heavy-tailed. In particular, if we choose a Gamma(γ, λ) prior for τ^{-2}, then the marginal prior is

$$\begin{aligned}\pi(\theta) &\propto \int (\tau^{-2})^{1/2} \exp\left(-\tau^{-2}\frac{(\theta-\mu)^2}{2}\right))(\tau^{-2})^{\gamma-1}\exp(-\lambda\tau^{-2})d(\tau^{-2}) \\ &= \int u^{\gamma-1/2}\exp\left(-(\lambda + \frac{(\theta-\mu)^2}{2})u\right)du \propto \frac{1}{\left(\lambda + \frac{(\theta-\mu)^2}{2}\right)^{\gamma+1/2}},\end{aligned}$$

where the last step follows by a change of the integration variable. This is a scaled and shifted t-density with 2γ degrees of freedom which has heavier tails than the normal distribution. The posterior does not belong to any standard family of distributions, but one can show that the posterior mode

$$\arg\max_{\theta} - \left(n(\theta-\bar{x})^2 + (2\gamma+1)\log(\lambda + \frac{(\theta-\mu)^2}{2})\right)$$

is close to $\bar{x}$ if prior and data are in conflict, that is, if $|\bar{x}-\mu|$ is large.

For the second approach, we already know $\pi(\theta \mid \tau^2, x_1, \dots, x_n)$. In Example 1.2 we also have found that for given τ^2, the marginal distribution of $(X_1, \dots, X_n)$ is normal with mean $\mu\mathbf{1}$ and covariance matrix $\Sigma = I + \tau^2\mathbf{1}\mathbf{1}^T$ where $\mathbf{1} = (1, \dots, 1)^T$. To simplify how $f(x_1, \dots, x_n \mid \tau^2)$ depends on τ^2, we consider a linear transformation $Y = AX$ where A is orthogonal and the first row is equal to $\frac{1}{\sqrt{n}}\mathbf{1}^T$. Then the Y_i are independent and the $Y_2, \dots, Y_n$ are standard normal for any value of μ and τ^2. Hence we obtain that

$$f(x_1, \dots, x_n \mid \tau^2) \propto (1+n\tau^2)^{-1/2}\exp\left(-\frac{n}{2(1+n\tau^2)}(\bar{x}-\mu)^2\right)$$

where $\propto$ means up to factors which do not depend on τ^2. This shows that the marginal posterior of τ^2 has little mass for small values if $(\bar{x}-\mu)^2$ is large, that is if prior mean and data are in conflict.

Example 3.2. *Hierarchical Poisson model. In insurance, the number of claims X_j of contract j during a given time period can be modeled as independent Poisson(θ_j) $(j = 1, 2, \dots, J)$ where the θ_j are i.i.d. and Gamma(γ, λ)-distributed (assuming the contracts have similar volumes). Then on the highest level we have the parameters (γ, λ), on the second level $(\theta_1, \dots, \theta_J)$ and on the lowest level $(X_1, \dots, X_J)$. The joint distribution factors as*

$$\pi(\gamma, \lambda)\prod_{j=1}^{J}\pi(\theta_j \mid \gamma, \lambda)f(x_j \mid \theta_j).$$

In the first approach, we would directly work with the prior

$$\pi(\theta_1,\ldots,\theta_J)=\int\frac{\lambda^{J\gamma}}{\Gamma(\gamma)^J}\prod_{j=1}^J\theta_j^{\gamma-1}\exp(-\lambda\sum\theta_j)\pi(\lambda,\gamma)d\lambda d\gamma.$$

Again this is not a standard distribution, and in particular the θ_j are not independent under this prior. They are only exchangeable, meaning that the density does not change if we permute the arguments. The second approach is somewhat simpler. First we observe that conditional on (γ,λ), the (θ_j,x_j) are independent for different j's and thus

$$\pi(\theta_1,\ldots,\theta_J\mid\gamma,\lambda,x_1,\ldots,x_J)=\prod_{j=1}^J\pi(\theta_j\mid\gamma,\lambda,x_j)$$

and

$$f(x_1,\ldots,x_J\mid\gamma,\lambda)=\prod_{j=1}^J f(x_j\mid\gamma,\lambda).$$

Moreover, by conjugacy

$$\pi(\theta_j\mid\gamma,\lambda,x_j)=\frac{(\lambda+1)^{\gamma+x_j}}{\Gamma(\gamma+x_j)}\theta_j^{\gamma+x_j-1}\exp(-(\lambda+1)\theta_j).$$

Because we know the normalizing constant for the Gamma density, we also obtain

$$f(x_j\mid\gamma,\lambda)=\int e^{-\theta_j}\frac{\theta_j^{x_j}}{x_j!}\theta_j^{\gamma-1}e^{-\lambda\theta_j}d\theta_j\frac{\lambda^\gamma}{\Gamma(\gamma)}=\frac{\Gamma(\gamma+x_j)}{x_j!\Gamma(\gamma)}\frac{\lambda^\gamma}{(\lambda+1)^{\gamma+x_j}}.$$

Finally, because $\Gamma(z+1)=z\Gamma(z)$, we end up with

$$f(x_j\mid\gamma,\lambda)=\frac{(\gamma+x_j-1)\cdots\gamma}{x_j\cdots 1}\left(\frac{\lambda}{\lambda+1}\right)^\gamma\left(\frac{1}{\lambda+1}\right)^{x_j}.$$

This is the negative binomial distribution with parameters γ and $p=\lambda/(\lambda+1)$. For integer γ it is the distribution of the number of failures until γ successes have been observed in a coin tossing experiment with success parameter p. These results can be used to compute

$$\begin{aligned}\pi(\theta_j\mid x_1,\ldots,x_n)&=\int\pi(\theta_j\mid\gamma,\lambda,x_j)\pi(\gamma,\lambda\mid x_1,\ldots,x_J)d\gamma d\lambda\\&\propto\int\pi(\theta_j\mid\gamma,\lambda,x_j)\prod_{i=1}^J f(x_i\mid\gamma,\lambda)\pi(\gamma,\lambda)d\gamma d\lambda.\end{aligned}$$

The last integral cannot be computed in closed form, but because it is only two-dimensional, approximations are possible. A simpler approach avoiding the computation of an integral will be discussed below in the section on empirical Bayes methods. The formula above implies that observations $x_i\neq x_j$ carry information about θ_j which seems somewhat counterintuitive: The insurance company uses the number of claims of other people to estimate the expected number of claims that I will have in a future period. This is justified by the assumption that I am just one member of a population and the company can estimate the distribution of the expected number of claims in this population from the number of claims of other members of this population.

Example 3.3. *The one-way ANOVA model. This model considers outcomes y_{ij} for different groups, indexed by i, each group consisting of a number of subjects, indexed by j:*

$$y_{ij} = \theta_i + \varepsilon_{ij} \quad (j = 1, \ldots, n_i; i = 1, \ldots, I),$$

where the $\varepsilon_{i,j}$ are i.i.d $\sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In order to be consistent with the notation in ANOVA models, we denote here the observations by y and not by x as in the previous examples. In the frequentist approach, we can interprete the group effects θ_i either as unobserved random variables (random effects) or as unknown parameters (fixed effects). The distinction would affect the way that the data are analyzed. In the Bayesian approach, we specify in both interpretations a distribution for $(\theta_1, \ldots, \theta_I)$. In this example one commonly assumes that the θ_i are i.i.d $\sim \mathcal{N}(\mu, \tau^2)$. In ANOVA terminology, μ is the grand mean and $\theta_i - \mu$ is the effect of group i. The hyperparameters are $\xi = (\mu, \tau^2)$, and we let μ and τ^2 be independent a priori. In order to simplify the formulae, we assume σ_ε^2 to be known. In Chapter 4, we will discuss how to proceed if also σ_ε^2 is unknown. The joint distribution of all unknown variables factors as

$$\pi(\mu)\pi(\tau^2) \prod_{i=1}^{I} \pi(\theta_i \mid \mu, \tau^2) \prod_{j=1}^{n_i} f(y_{ij} \mid \theta_i).$$

The priors for μ and τ^2 will be chosen later, the other factors are normal densities.

We will proceed according to the second approach. First we observe that

$$\prod_{j=1}^{n_i} f(y_{ij} \mid \theta_i) = f(\bar{y}_{i.} \mid \theta_i) f(y_{i1}, \ldots, y_{i,n_i-1} \mid \bar{y}_{i.}),$$

where $\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i$ denotes the mean in group i. That the second factor on the right does not depend on θ_i follows by a similar argument to the one given in Example 3.1. As a consequence we only need to consider the first factor for computing posterior distributions.

By an argument that should be familiar by now we can show that conditional on (y, μ, τ^2) the θ_i are independent and

$$\theta_i \mid y, \mu, \tau^2 \sim \mathcal{N}\left(\frac{n_i \tau^2 \bar{y}_{i.} + \sigma_\varepsilon^2 \mu}{n_i \tau^2 + \sigma_\varepsilon^2}, \frac{n_i \tau^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2 \tau^2}\right).$$

In order to compute the posterior of the hyperparameters, $\pi(\mu, \tau^2 \mid y)$, we use

$$\pi(\mu, \tau^2 \mid y) = \pi(\mu \mid \tau^2, y)\pi(\tau \mid y) = \pi(\mu \mid \tau^2, (\bar{y}_{i.}))\pi(\tau \mid (\bar{y}_{i.}))$$

and

$$\pi(\mu \mid \tau^2, (\bar{y}_{i.})) \propto \pi(\mu) \prod_{i=1}^{I} f(\bar{y}_{i.} \mid \mu, \tau^2).$$

Because $\bar{y}_{i.} = \theta_i + \bar{\varepsilon}_{i.}$, it follows that conditionally on μ and τ^2, $\bar{y}_{i.}$ is normal with mean μ and variance $\tau^2 + \sigma_\varepsilon^2/n_i$. Hence we can choose the improper flat prior $\pi(\mu) = 1$ and obtain – again by a standard argument – the proper posterior

$$\mu \mid \tau^2, (\bar{y}_{i.}) \sim \mathcal{N}(\widehat{\mu}, V_\mu), \quad \widehat{\mu} = \frac{\sum_{i=1}^{I} w_i \bar{y}_{i.}}{\sum_{i=1}^{I} w_i}, V_\mu = \frac{1}{\sum_{i=1}^{I} w_i}, w_i = \frac{n_i}{n_i \tau^2 + \sigma_\varepsilon^2}.$$

Finally, we need to compute $\pi(\tau^2 \mid (\bar{y}_{i.}))$. For this we can either integrate μ out:

$$\pi(\tau^2 \mid (\bar{y}_{i.})) \propto \int \pi(\mu)\pi(\tau^2) \prod_{i=1}^{I} f(\bar{y}_{i.} \mid \mu, \tau^2) d\mu$$

or use the identity

$$\pi(\tau^2 \mid (\bar{y}_{i.})) = \frac{\pi(\mu, \tau^2 \mid (\bar{y}_{i.}))}{\pi(\mu \mid \tau^2, (\bar{y}_{i.}))} \propto \frac{\pi(\mu)\pi(\tau^2)\prod_{i=1}^{I} f(\bar{y}_{i.} \mid \mu, \tau^2)}{\pi(\mu \mid \tau^2, (\bar{y}_{i.}))}$$

which holds for an arbitrary value of μ. If we choose $\mu = \widehat{\mu}$, we obtain

$$\pi(\tau \mid (\bar{y}_{i.})) \propto \pi(\tau^2) V_\mu^{1/2} \prod_{i=1}^{I} \left(\frac{n_i}{\sigma_\varepsilon^2 + n_i \tau^2} \right)^{1/2} \exp\left(-\frac{n_i}{\sigma_\varepsilon^2 + n_i \tau^2} (\bar{y}_{i.} - \widehat{\mu})^2 \right).$$

This is not a standard distribution. As $\tau^2 \to 0$, the terms on the right without the prior $\pi(\tau^2)$ have a finite, strictly positive limit. This implies that for a proper posterior we should choose a prior for τ^2 which is integrable at the origin. This rules out Jeffreys prior, we need some prior information about small values of τ^2.

The main interest is in the posterior distribution of θ_i given y which is given by the formula

$$\pi(\theta_i \mid y) = \int \int \pi(\theta_i \mid y, \mu, \tau^2) \pi(\mu \mid y, \tau^2) d\mu \; \pi(\tau^2 \mid y) d\tau^2.$$

The inner integral can be computed in closed form. It gives a normal density with mean equal to a convex combination of all group means $\bar{y}_{1.}, \ldots \bar{y}_{I.}$. The weights in the comvex combination and the variance depend on τ^2, and the integration over τ^2 has to be done by one of the approximate methods that we will discuss in Chapter 4.

3.2 Empirical Bayes methods

In the examples of the preceeding section we have seen that in a hierarchical Bayes model with conjugate priors, we can compute $\pi(\theta \mid x, \xi)$ and $f(x \mid \xi)$ in closed form, but then we need to use approximations to compute

$$\pi(\theta \mid x) \propto \int \pi(\theta \mid x, \xi) f(x \mid \xi) \pi(\xi) d\xi.$$

The empirical Bayes method uses instead

$$\pi(\theta \mid x) \approx \pi(\theta \mid x, \widehat{\xi}(x)), \quad \widehat{\xi}(x) = \arg\max_\xi f(x \mid \xi).$$

This means instead of taking a weighted average, we take the value with maximal weight (assuming $\pi(\xi)$ is flat around $\widehat{\xi}(x)$). $\widehat{\xi}(x)$ is simply the marginal maximum likelihood estimator of the hyperparameter.

This method avoids not only the computation of the integral, but also the choice of a hyperprior $\pi(\xi)$. From a conceptual point of view, it is however not quite satisfactory because it uses the data x twice: First to select the prior $\pi(\theta \mid \widehat{\xi}(x))$ and then to compute the posterior according to Bayes formula. It is also clear that in general the uncertainty is underestimated if we choose any particular value of ξ instead of averaging over different plausible values. For these reasons, Bayesians try to avoid empirical Bayes methods as far as possible, and in fact empirical Bayes methods are usually justified by their frequentist properties. But from a pragmatic point of view, empirical Bayes methods are useful and they offer a way to exit from the infinite hierarchy that arises if we acknowledge that the hyperprior $\pi(\xi)$ is also uncertain and thus we should introduce another prior to describe it, etc.

Let us look what empirical Bayes estimation gives for the examples of the previous section.

Example 3.4. *Normal means, ctd. In example 3.1, we have already computed $f(x_1, \dots, x_n \mid \tau^2)$. Maximizing the result with respect to τ^2 is straightworward. We obtain*

$$\widehat{\tau}^2 = max(0, (\bar{x} - \mu)^2 - 1/n).$$

This means that if the prior mean μ conflicts with the data, we choose a wider prior. In particular, we have

$$\mathbb{E}(\theta \mid x_1, \dots, x_n) = \mu \quad (|\bar{x} - \mu| \leq 1/\sqrt{n}), \qquad = \bar{x} + \frac{1}{n(\mu - \bar{x})} \quad (|\bar{x} - \mu| \geq 1/\sqrt{n})$$

which has the asymptote $\bar{x}$ as $|\bar{x}| \to \infty$. However, the posterior variance is zero for $|\bar{x} - \mu| \leq 1/\sqrt{n}$ which is not reasonable, confirming the intuition that empirical Bayes underestimates the uncertainty.

Example 3.5. *Hierarchical Poisson model, ctd. Maximizing the marginal likelihood is equivalent to minimizing*

$$\gamma \log(1 + 1/\lambda) + \sum_{j=1}^{n} x_j \log(1 + \lambda) + \sum_{j; x_j > 0} \sum_{k=0}^{x_j - 1} \log(\gamma + k)$$

with respect to γ and λ. Setting the partial derivative with respect to λ equal to zero gives

$$\frac{\widehat{\gamma}}{\widehat{\lambda}} = \frac{1}{n} \sum_{j=1}^{n} x_j = \bar{x},$$

but setting the partial derivative with respect to γ equal to zero gives an equation that does not have an explicit solution. If one does not want to find the minimum numerically, one can use instead of the marginal MLE the marginal moment estimator. By standard rules from probability theory and results about the moments of the Gamma distribution, we obtain

$$\mathbb{E}(X_j \mid \gamma, \lambda) = \mathbb{E}(\mathbb{E}(X_j \mid \theta_j) \mid \gamma, \lambda) = \mathbb{E}(\theta_j \mid \gamma, \lambda) = \frac{\gamma}{\lambda}$$

and

$$\mathrm{Var}(X_j \mid \gamma, \lambda) = \mathrm{Var}(\mathbb{E}(X_j \mid \theta_j) \mid \gamma, \lambda) + \mathbb{E}(\mathrm{Var}(X_j \mid \theta_j) \mid \gamma, \lambda) = \frac{\gamma}{\lambda^2} + \frac{\gamma}{\lambda} = \frac{\gamma(\lambda + 1)}{\lambda^2}.$$

(alternatively, we could use known results for the moments of the negative binomial distribution). Therefore the marginal moment estimator is

$$\widehat{\lambda} = \frac{\bar{x}}{S_n^2 - \bar{x}}, \quad \widehat{\gamma} = \bar{x}\widehat{\lambda}$$

where S_n is the sample variance $(n-1)^{-1} \sum_i (x_i - \bar{x})^2$. The empirical Bayes estimate of θ_j is therefore

$$\mathbb{E}(\theta_j \mid \widehat{\lambda}, \widehat{\gamma}) = \frac{\widehat{\lambda}}{\widehat{\lambda} + n}\, \bar{x} + \frac{n}{\widehat{\lambda} + n}\, x_j.$$

It shrinks the individual experience x_j towards the average experience of all contracts. Because $\widehat{\lambda}$ is a decreasing function of S_n^2, the shrinkage is less if the portfolio of all contracts is heterogeneous. The method breaks down if $S_n^2 \leq \bar{x}$.

Example 3.6. *The one-way ANOVA model, ctd. In order to simplify the formulae, we assume that $\sigma_\varepsilon^2 = 1$ and $n_i = 1$ for all i. Then we need only a single index i. Because conditionally on μ and τ^2, $y_i = \theta_i + \varepsilon_i$ is normal with mean μ and variance $\tau^2 + 1$, the marginal MLE is easily seen to be*

$$\widehat{\mu} = \bar{y}, \quad \widehat{\tau}^2 = \frac{1}{I}\sum_{i=1}^{I}(y_i - \bar{y})^2 - 1.$$

Therefore the empirical Bayes estimator of θ_i is

$$\mathbb{E}(\theta_i \mid \widehat{\mu}, \widehat{\tau}^2) = \frac{I}{\sum_{i=1}^{I}(y_i - \bar{y})^2}\, \bar{y} + \left(1 - \frac{I}{\sum_{i=1}^{I}(y_i - \bar{y})^2}\right) y_i.$$

Again we have a shrinkage of the individual observation towards the mean of all observations, and the amount of shrinkage depends on how heterogeneous the sample is. As in the previous example, there are problems when $\sum_{i=1}^{I}(y_i - \bar{y})^2 < 1$ because y_i has then a negative weight.

A famous result of Charles Stein says that if we use the loss function $L(t,\theta) = \sum_i (t_i - \theta_i)^2$ and if $I > 3$, then this empirical Bayes estimator has for any θ a smaller risk than the unbiased estimator $\widehat{\theta}_i = y_i$. This is a frequentist justification for an empirical Bayes method.

3.3 Model selection in linear regression

We consider here the linear regression model

$$y = \alpha\mathbf{1} + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Here y is an $n \times 1$ vector of responses, $\mathbf{1} = (1, \ldots, 1)^T$, X is the $n \times p$ design matrix, α is the intercept, β the $p \times 1$ regression parameter and ε the $n \times 1$ vector of errors. The j-th column of X contains the values of the j-th explanatory variable, and we assume that all columns are centered: $X^T\mathbf{1} = 0$. We consider not only estimation of the parameters, but also the selection of explanatory variables. Let γ denote an element of $\{0,1\}^p$ where $\gamma_j = 1$ iff the j-th variable is selected. Furthermore, let β_γ be the subvector that contain only the selected components and X_γ the corresponding submatrix. Then the model indexed by γ is

$$y = \alpha\mathbf{1} + X_\gamma\beta_\gamma + \varepsilon.$$

The number of explanatory variables included in the model γ will be denoted by $|\gamma|$. The unknowns are $(\gamma, \beta_\gamma, \alpha, \sigma^2)$ and we would like to compute the posterior of these unknowns. For this we need the likelihood and a prior.

3.3.1 g-Prior and posterior for fixed g and γ

The likelihood is

$$(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - \alpha\mathbf{1} - X_\gamma\beta_\gamma)^T(y - \alpha\mathbf{1} - X_\gamma\beta_\gamma)\right).$$

Because the columns of X are assumed to be centered, the MLE for α and β_γ is

$$\widehat{\alpha} = \bar{y}, \quad \widehat{\beta}_\gamma = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y,$$

and the likelihood is equal to

$$\begin{aligned}&(\sigma^2)^{-n/2}\exp\left(-\frac{(y-\widehat{\alpha}\mathbf{1}-X_\gamma\widehat{\beta}_\gamma)^T(y-\widehat{\alpha}\mathbf{1}-X_\gamma\widehat{\beta}_\gamma)+n(\widehat{\alpha}-\alpha)^2+(\beta_\gamma-\widehat{\beta}_\gamma)X_\gamma^TX_\gamma(\beta_\gamma-\widehat{\beta}_\gamma)}{2\sigma^2}\right)\\ =\;&(\sigma^2)^{-n/2}\exp\left(-\frac{s_\gamma^2+n(\widehat{\alpha}-\alpha)^2+(\beta_\gamma-\widehat{\beta}_\gamma)X_\gamma^TX_\gamma(\beta_\gamma-\widehat{\beta}_\gamma)}{2\sigma^2}\right)\end{aligned}$$

where

$$s_\gamma^2=(y-\widehat{\alpha}\mathbf{1}-X_\gamma\widehat{\beta}_\gamma)^T(y-\widehat{\alpha}\mathbf{1}-X_\gamma\widehat{\beta}_\gamma)$$

is the residual sum of squares in the model defined by γ.

In this subsection, we fix the model γ. We assume α to be independent a priori from σ^2 and β_γ, and we take univariate Jeffreys priors for α and σ^2. This means

$$\pi(\beta_\gamma,\alpha,\sigma^2)=\pi(\alpha)\pi(\sigma^2)\pi(\beta_\gamma\mid\sigma^2)\propto\pi(\beta_\gamma\mid\sigma^2)\sigma^{-2}.$$

For the first factor, a popular choice is the so-called g-prior of Zellner

$$\beta_\gamma\mid\sigma^2\sim\mathcal{N}(0,g\sigma^2(X_\gamma^TX_\gamma)^{-1}).$$

Because in regression models the design matrix is considered to be known and fixed, it is allowed to use it for the prior. The dependence on σ^2 is necessary for conjugacy. $g>0$ is a hyperparameter which describes the uncertainty of the prior mean 0. More specifically, the above formula for the likelihood shows that this prior arises as the posterior from a flat prior and a response vector zero with the same design matrix X_γ, no intercept and error variance $g\sigma^2$. For g large, this prior is therefore weakly informative, and for $g\to\infty$, we approach a flat prior. But we cannot use a flat prior if we want to do model selection later because this would leave posterior probabilities of different models γ undefined (see below).

Combining this prior with the likelihood above leads to the posterior

$$\begin{aligned}\pi(\beta_\gamma,\alpha,\sigma^2\mid y)\;&\propto\;(\sigma^2)^{-n/2-1}\exp\left(-\frac{s_\gamma^2}{2\sigma^2}\right)(g\sigma^2)^{|\gamma|/2}\det(X_\gamma^TX_\gamma)^{1/2}\\ &\cdot\;\exp\left(-\frac{n(\widehat{\alpha}-\alpha)^2+(\beta_\gamma-\widehat{\beta}_\gamma)^TX_\gamma^TX_\gamma(\beta_\gamma-\widehat{\beta}_\gamma)+\frac{1}{g}\beta_\gamma^TX_\gamma^TX_\gamma\beta_\gamma}{2\sigma^2}\right).\end{aligned}$$

As in previous examples involving the normal distribution, we complete the square

$$\begin{aligned}&\left(\beta_\gamma-\widehat{\beta}_\gamma\right)^TX_\gamma^TX_\gamma\left(\beta_\gamma-\widehat{\beta}_\gamma\right)+\frac{1}{g}\beta_\gamma^TX_\gamma^TX_\gamma\beta_\gamma\\ &=\;\frac{g+1}{g}(\beta_\gamma-\frac{g}{g+1}\widehat{\beta}_\gamma)^TX_\gamma^TX_\gamma(\beta_\gamma-\frac{g}{g+1}\widehat{\beta}_\gamma)+\frac{1}{g+1}\widehat{\beta}_\gamma^TX_\gamma^TX_\gamma\widehat{\beta}_\gamma.\end{aligned}$$

Then the posterior becomes

$$\begin{aligned}\pi(\beta_\gamma,\alpha,\sigma^2\mid y)\;&\propto\;\left(\frac{n}{\sigma^2}\right)^{1/2}\exp\left(-\frac{n}{2\sigma^2}(\widehat{\alpha}-\alpha)^2\right)\cdot\left(\frac{g+1}{g\sigma^2}\right)^{|\gamma|/2}\det(X_\gamma^TX_\gamma)^{1/2}\\ &\cdot\;\exp\left(-\frac{g+1}{2g\sigma^2}\left(\beta_\gamma-\frac{g}{g+1}\widehat{\beta}_\gamma\right)^TX_\gamma^TX_\gamma\left(\beta_\gamma-\frac{g}{g+1}\widehat{\beta}_\gamma\right)\right)\\ &\cdot\;(\sigma^2)^{-(n+1)/2}(g+1)^{-|\gamma|/2}\exp\left(-\frac{s_\gamma^2+\frac{1}{g+1}\widehat{\beta}_\gamma^TX_\gamma^TX_\gamma\widehat{\beta}_\gamma}{2\sigma^2}\right).\end{aligned}$$

From this we conclude that

$$\beta_\gamma \mid y, \sigma^2 \sim \mathcal{N}\left(\frac{g}{g+1}\widehat{\beta}_\gamma, \frac{g\sigma^2}{g+1}(X_\gamma^T X_\gamma)^{-1}\right)$$

and

$$\alpha \mid y, \sigma^2 \sim N(\bar{y}, \frac{\sigma^2}{n}),$$

independently of β_γ. Furthermore, integrating with respect to α and β_γ gives us

$$\pi(\sigma^2 \mid y) \propto (\sigma^2)^{-(n-|\gamma|-1)/2-1} \exp\left(-\frac{1}{2\sigma^2}(s_\gamma^2 + \frac{1}{g+1}\widehat{\beta}_\gamma^T X_\gamma^T X_\gamma \widehat{\beta}_\gamma)\right).$$

This means that

$$\sigma^{-2} \mid y \sim \text{Gamma}\left(\frac{n-|\gamma|-1}{2}, \frac{s_\gamma^2 + \frac{1}{g+1}\widehat{\beta}_\gamma^T X_\gamma^T X_\gamma \widehat{\beta}_\gamma}{2}\right).$$

By the orthogonality properties of least squares, it follows that

$$\widehat{\beta}_\gamma^T X_\gamma^T X_\gamma \widehat{\beta}_\gamma = s_0^2 - s_\gamma^2$$

where $s_0^2 = (y - \bar{y}\mathbf{1})^T(y - \bar{y}\mathbf{1})$ is the sum of squared errors in the null model $\gamma = 0$. Therefore

$$s_\gamma^2 + \frac{1}{g+1}\widehat{\beta}_\gamma^T X_\gamma^T X_\gamma \widehat{\beta}_\gamma = \frac{s_0^2(1 + g(1 - R_\gamma^2))}{g+1}$$

where

$$R_\gamma^2 = 1 - \frac{s_\gamma^2}{s_0^2}.$$

is the so-called coefficient of determination of model γ.

3.3.2 Model selection

For Bayesian model selection, we put a prior on the set of possible models γ and compute the posterior

$$\pi(\gamma \mid y) = \frac{\pi(\gamma) f(y \mid \gamma)}{\sum_{\gamma'} \pi(\gamma') f(y \mid \gamma')}$$

where the marginal likelihood $f(y \mid \gamma)$ is

$$f(y \mid \gamma) = \int f(y \mid \beta_\gamma, \alpha, \sigma^2)\pi(\beta_\gamma, \alpha, \sigma^2) d\beta_\gamma d\alpha d\sigma^2.$$

If we had chosen an improper prior for β_γ, posterior model probabilities would not be well defined even if the posterior $\pi(\beta_\gamma \mid y)$ is well defined: An improper prior means that $f(y \mid \gamma)$ is only defined up to an arbitrary constant which does *not* cancel in $\pi(\gamma \mid y)$ because this constant differs for different models. An improper prior for α and σ^2 is allowed because these two parameters are shared by all models.

For the g-prior, $f(y \mid \gamma)$ can be computed in closed form by the same arguments that were used to compute the posterior $\pi(\beta_\gamma, \alpha, \sigma^2 \mid y)$. We skip the details and just state the result:

$$f(y \mid \gamma) \propto \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1 + g(1 - R_\gamma^2))^{(n-1)/2}}$$

where $\propto$ means up to factors which contain neither γ nor g.

The posterior model probabilities depend also on the prior. Bayesian model comparison is usually based on the Bayes factor, the ratio of posterior and prior odds for two models, which is independent of the prior:

$$B(\gamma,\gamma') = \frac{\pi(\gamma \mid y)}{\pi(\gamma' \mid y)} \frac{\pi(\gamma')}{\pi(\gamma)} = \frac{f(y \mid \gamma)}{f(y \mid \gamma')} = (1+g)^{(|\gamma'|-|\gamma|)/2} \left(\frac{1+g(1-R^2_{\gamma'})}{1+g(1-R^2_{\gamma})} \right)^{(n-1)/2}.$$

In particular, the Bayes factor for comparing with the null model is

$$B(\gamma,0) = \frac{(1+g)^{(n-|\gamma|-1)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}}.$$

The second factor compares how well the two models fit the data: If γ' is a sub-model of γ, that is every variable included in γ' is also included in γ, then by the definition of least squares $s^2_{\gamma'} \geq s^2_\gamma$ and $R^2_{\gamma'} \leq R^2_\gamma$. The second factor is therefore greater (or equal) to one. However, the first factor is a penalty for the more complex model (measured by the number of parameters). In particular it is less than one if γ' is a sub-model of γ. Hence the Bayes factor balances goodness of fit and complexity when comparing two models.

Clearly, the Bayes factors depend strongly on the chosen value of g. In an attempt to be non informative, one is tempted to let g tend to infinity. However, as $g \to \infty$, $B(\gamma,0) \to 0$ for any $\gamma \neq 0$, that is we always choose the null model ("Bartlett's paradox"). Choosing any fixed value for g also leads to problems: If one model γ has an excellent fit, one would expect that the Bayes factor clearly favors this model over the null model. However, if g is fixed and $R^2_\gamma \to 1$ then $B(\gamma,0) \to (1+g)^{(n-1-\gamma)/2}$ which is finite although the evidence in favor of γ as measured for instance by the p value of the F-test goes to infinity ("Information paradox").

For prediction of a new observation y_{n+1} for a given vector x_{n+1} of explanatory variables, a Bayesian prefers to use model averaging instead of model selection. For instance, the prediction of the mean of y_{n+1} is (for known g) given by

$$\mathbb{E}(y_{n+1} \mid y) = \bar{y} + \frac{g}{g+1} \sum_\gamma x^T_{n+1,\gamma} \widehat{\beta}_\gamma \pi(\gamma \mid y).$$

If we want to use posterior model probabilities instead of Bayes factors, we also have to choose a prior for $\gamma \in \{0,1\}^p$. The simplest choice is the uniform prior $\pi(\gamma) = 2^{-p}$ for all γ which means that each explanatory variable is included with probability $\frac{1}{2}$, independently of the others. For large p his is however informative for the size of the model because with high prior probability $|\gamma| \approx \frac{p}{2}$. In order to have an uniform prior for $|\gamma|$ one can assume that each explanatory variable is included with probability r where r is unknown and uniform on $(0,1)$.

If the number p of variables is large, then computing $\pi(\gamma \mid y)$ for all γ and finding the a posteriori most likely model is difficult. In such a situation, stochastic search algorithms which avoid enumerating the whole space $\{0,1\}^p$ systematically are being recommended.

3.3.3 Unknown g

In order to avoid the paradoxes described in the previous section, we should consider g to be unknown. We can then either use an empirical Bayes approach or a fully Bayesian approach with a hyperprior on g.

In the empirical Bayes approach, we can determine $\widehat{g}$ either separately for each model γ, or globablly. The former means

$$\widehat{g} = \arg\max((n-1-|\gamma|)\log(1+g)-(n-1)\log(1+g(1-R_\gamma^2))) = \max\left(\frac{(n-1-|\gamma|)R_\gamma^2}{|\gamma|(1-R_\gamma^2)} - 1, 0\right).$$

The ratio on the right is the standard F-test statistics for the null hypothesis $\beta_\gamma = 0$. A large values of this test statistic means that the data are in conflict with the prior mean zero of β_γ and therefore the influence of the prior is reduced. In the latter case, we have

$$\widehat{g} = \arg\max \sum_\gamma \pi(\gamma) f(y \mid \gamma)$$

which has to be computed numerically. In both cases one can show that the information paradox does not occur any more.

In the fully Bayesian approach it is desirable to have a prior $\pi(g)$ such that

$$f(y \mid \gamma) \propto \int \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1+g(1-R_\gamma^2))^{(n-1)/2}} \pi(g) dg$$

can be computed easily. Moreover, in order to avoid the information paradox, one should have

$$\int (1+g)^{(n-1-|\gamma|)/2} \pi(g) dg = \infty \quad (|\gamma| \leq p).$$

In the case of the minimal sample size $n = p + 2$, this means $\int (1+g)^{1/2}\pi(g)dg = \infty$. In the literature, the choices

$$\pi(g) \propto g^{-3/2} \exp(-n/(2g))$$

and

$$\pi(g) \propto (1+g)^{-a/2}, \quad (a < 2 \leq 3)$$

have been proposed. With the second choice one can express $f(y \mid \gamma)$ with the so-called Gaussian hypergeometric function (Abramowitz and Stegun, 1970, Chapter 15). Moreover, the shrinkage factor $\mathbb{E}(\frac{g}{g+1} \mid y, \gamma)$ which appears in $\mathbb{E}(\beta_\gamma \mid y, \gamma)$ can also be expressed with the same Gaussian hypergeometric function.

For more details on Bayesian model selection in regression, see Feng et al., J. Amer. Statist. Assoc. 103, 2008, p. 410.

Chapter 4

Bayesian computation

Until now we have mainly worked with conjugate priors where the posterior is a standard distribution and moments or quantiles are either known explicitly or can be expressed by functions that are coded efficiently in R (or other software packages). But in the case of hierarchical models we already ended up with distributions that are not standard and contain integrals that cannot be computed analytically. Most applications today involve hierarchical models with many parameters, and often on the lowest level likelihoods occur for which conjugate priors do not exist. Thus the posterior is usually high-dimensional with complex dependencies and does not belong to a standard family of distributions. In order to compute moments, quantiles, marginal densities or posterior predictive distributions, we need approximate methods to compute integrals.

4.1 Laplace approximation

Laplace approximations are used for integrals of the form

$$\int h(\theta)q(\theta)d\theta$$

where q is a possibly unnormalized smooth density which is concentrated around its mode $\theta_0 = \arg\max \log q(\theta)$ and where h is an arbitrary smooth function. Because the gradient of $\log q$ at θ_0 is zero, we have

$$\log q(\theta) \approx \log q(\theta_0) - \frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)$$

where J is minus the Hessian matrix with elements

$$J(\theta)_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log q(\theta).$$

Because q is maximal at θ_0, $J(\theta_0)$ is positive definite. Expanding h into a first order Taylor series at θ_0, we obtain from the formula for the normalizing constant of a normal density

$$\begin{aligned}
\int h(\theta)q(\theta)d\theta &\approx h(\theta_0)q(\theta_0)\int \exp\left(-\frac{1}{2}(\theta-\theta_0)^T J(\theta_0)(\theta-\theta_0)\right) d\theta \\
&+ q(\theta_0)\frac{\partial h}{\partial\theta}(\theta_0)^T \int (\theta-\theta_0)\exp\left(-\frac{1}{2}(\theta-\theta_0)^T J(\theta_0)(\theta-\theta_0)\right) d\theta \\
&= h(\theta_0)q(\theta_0)(\det J(\theta_0))^{-1/2}(2\pi)^{p/2}.
\end{aligned}$$

If we want to approximate a posterior expectation

$$\mathbb{E}(h(\theta) \mid x) = \frac{\int h(\theta)\pi(\theta)f(x \mid \theta)d\theta}{\int \pi(\theta)f(x \mid \theta)d\theta}$$

we can use the Laplace approximation for the numerator and the denominator separately and take either $q(\theta) = f(x \mid \theta)$ or $q(\theta) = \pi(x)f(x \mid \theta)$. If h is strictly positive, we can also take $q(\theta) = h(\theta)\pi(\theta)f(x \mid \theta)$ in the numerator. All these variants lead to slightly different approximations.

It is clear from the derivation that the Laplace approximation is good if q has a sharp peak at θ_0 and is much smaller everywhere outside of a neighborhood of θ_0. For a rigorous statement of a limit theorem justifying the use of the Laplace approximation, one has to consider a sequence of integrands that get sharper and sharper, that is $J(\theta_0) \to \infty$. This occurs for instance if we consider the posterior in an i.i.d. model with n observations where

$$q_n(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta),$$

θ_0 is the MLE $\widehat{\theta}_n$ and

$$J_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log f(x_i \mid \theta)$$

is the observed Fisher information. By the law of large numbers

$$J_n(\theta) \approx nI(\theta).$$

Example 4.1. . *Bayes factors and BIC. Consider two different models M_1 and M_2 for i.i.d data $X_j \sim f_i(x \mid \theta_i)dx$ with parameters $\theta_i \in \mathbb{R}^{p_i}$ and priors π_i $(i = 1, 2)$. Then the Bayes factor of model 1 with respect to model 2 is*

$$B_{12}(x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid M_1)}{f(x_1, \ldots, x_n \mid M_2)}, \quad f(x_1, \ldots, x_n \mid M_i) = \int \pi_i(\theta_i) \prod_{j=1}^{n} f_i(x_j \mid \theta_i)d\theta_i.$$

The Laplace expansion gives

$$f(x_1, \ldots, x_n \mid M_i) \approx \pi_i(\widehat{\theta}_i) \prod_{j=1}^{n} f_i(x_j \mid \widehat{\theta}_i)(\det(nI_i(\widehat{\theta}_i)))^{-1/2}(2\pi)^{p_i/2}$$

where $\widehat{\theta}_i$ is the MLE and $I_i(\theta)$ is the Fisher information in Model M_i. By the rules for the determinant, it follows that

$$\log f(x_1, \ldots, x_n \mid M_i) \approx \sum_{j=1}^{n} \log f_i(x_j \mid \widehat{\theta}_i) - \frac{p_i}{2} \log n + O(1).$$

The first term on the right measures how well the model fits the data and the second term is a penalty for model complexity. The two terms together can thus be used as a model selection criterion, the so-called the Bayesian information criterion (BIC). The Bayes factor is then the difference between the BIC values of the two models

Various extensions of the Laplace approximation are possible: If $q(\theta)$ is bimodal with modes at θ_0 and θ_1 that are well separated, we can use approximations around θ_0 and θ_1 separately. We can also use higher order Taylor approximations. To simplify the notation, assume that θ is one-dimensional. Then a fourth order Taylor approximation of $\log q$ together with $e^x \approx 1 + x$ gives

$$q(\theta) \approx q(\theta_0) \exp\left(-\frac{J(\theta_0)}{2}(\theta-\theta_0)^2\right)\left(1 + \frac{d^3 \log q(\theta_0)}{d\theta^3}\frac{(\theta-\theta_0)^3}{6} + \frac{d^4 \log q(\theta_0)}{d\theta^4}\frac{(\theta-\theta_0)^4}{24}\right).$$

The resulting integral can then be computed using results about higher moments of normal random variables.

4.2 Independent Monte Carlo methods

A Monte Carlo algorithm draws samples from a target distribution π on the basis of a sequence of i.i.d. uniform random variables (U^t). In applications, one uses pseudo-random numbers generated by a deterministic algorithm instead of truly random uniform sequences. How to get such pseudo-random numbers is a topic in itself, but we don't discuss this issue and just rely on the numbers provided by the software (e.g. by *R*). We assume that these pseudo-random numbers are good enough so that our results are not affected by this lack of true randomness.

Assume $(X^1, \ldots X^N)$ is an i.i.d. sample from π. We use superscripts to identify the members of the sample because often π is a distribution on $\mathbb{R}^p$ for $p > 1$ and we use subscripts to indicate the different components of X. Also in Bayesian applications π is the posterior distribution, so the variable is θ and not X. But Monte Carlo methods are also used outside Bayesian statistics, and thus we use here the more familiar notation X for random variables. By the law of large numbers we can estimate the expectation of an arbitrary function $h(X)$ by the sample average

$$\int h(x)\pi(x)dx = \mathbb{E}_\pi(h(X)) \approx \bar{H}_N := \frac{1}{N}\sum_{t=1}^N h(X^t).$$

Moreover, by the Central Limit Theorem

$$\frac{\sqrt{N}(\bar{H}_N - \int h(x)\pi(x)dx)}{\sqrt{\sum_{t=1}^N (h(X^t) - \bar{H}_N)^2/N}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1).$$

This can be used to compute a (frequentist) confidence interval for $\int h(x)\pi(x)dx$. The convergence rate $N^{-1/2}$ is rather slow, but the advantage is that – in contrast to numerical approximations – it is independent of the dimension and doesn't require h to be smooth (only $\int h(x)^2\pi(x)dx < \infty$ is needed). These are essential advantages in Bayesian computations.

If π is one-dimensional and we can compute the quantile function

$$F_\pi^{-1}(u) = \inf\{x; F_\pi(x) \geq u\}, \quad F_\pi(x) = \int_{-\infty}^x \pi(x')dx',$$

then $X^t = F_\pi^{-1}(U^t) \sim \pi$. In higher dimensions we can use the method in principle sequentially to draw first X_1 from the first marginal, then X_2 from the conditional distribution of the second component given X_1 , etc. In Bayesian statistics this is rarely feasible because the target π is known only up to a normalizing constant, and marginal or conditional quantiles cannot be computed explicitly.

There are two methods which both start by drawing a sample from some other distribution, the so-called proposal τ, and then convert it to a sample from the target π. The accept/reject methods assumes that the ratio of densities is bounded:

$$\frac{\pi(x)}{\tau(x)} \leq M < \infty.$$

If we are given two independent i.i.d. sequences (Y^t) and (U^t) where $Y^t \sim \tau$ and $U^t \sim$ uniform(0,1), then the following algorithm produces an i.i.d. sequence (X^t) where $X^t \sim \pi$: To start, set $t = 1$ and $s = 1$ and repeat the following

- If $U^s \leq \frac{\pi(Y^s)}{M\tau(Y^s)}$ set $X^t = Y^s$, $t = t + 1$, $s = s + 1$ (i.e. accept Y^s);
- else set $s = s + 1$ whereas t is unchanged (i.e. reject Y^s).

To see that this algorithm is correct, we use the heuristic interpretation $\mathbb{P}(Y \in dx) = \tau(x)dx$ of the density τ. Then by Bayes formula

$$\begin{aligned}\mathbb{P}(X \in dx) &= \mathbb{P}\left(Y \in dx \mid U \leq \frac{\pi(Y)}{M\tau(Y)}\right) \propto \mathbb{P}\left(U \leq \frac{\pi(Y)}{M\tau(Y)} \mid Y \in dx\right) \mathbb{P}(Y \in dx) \\ &= \frac{\pi(x)\tau(x)dx}{M\tau(x)} \propto \pi(x)dx.\end{aligned}$$

The algorithm works for any dimension, and τ and π need to be known only up to normalizing constants. But in practice, the expected number of rejections is large unless τ is reasonably close to π, and in high dimensions it is usually difficult to find a proposal which is close to the the target and from which we can simulate.

The second method is called importance sampling. It corrects by weighting the sample values $Y^t \sim \tau$:

$$\int h(x)\pi(x)dx \approx \frac{\sum_{t=1}^N h(Y^t)w(Y^t)}{\sum_{t=1}^N w(Y^t)}, \quad w(x) \propto \frac{\pi(x)}{\tau(x)}.$$

By the law of large numbers applied to the denominator and the numerator it follows that almost surely

$$\frac{\sum_{t=1}^N h(Y^t)w(Y^t)}{\sum_{t=1}^N w(Y^t)} \to \frac{\int h(x)w(x)\tau(x)dx}{\int w(x)\tau(x)dx} = \frac{\text{const.} \int h(x)\pi(x)dx}{\text{const.} \int \pi(x)dx} = \int h(x)\pi(x)dx.$$

For the first equality, we need that $\pi(x) > 0$ implies $\tau(x) > 0$. One can also show that a central limit theorem holds where the limiting variance is

$$\frac{\int w(x)^2(h(x) - \int h(x')\pi(x')dx')^2\tau(x)d(x)}{(\int w(x)\tau(x)dx)^2}.$$

This asymptotic variance can easily be estimated from the sample. As with accept/reject, importance sampling can be used in any dimension and the proposal and the target need to be known only up to a normalizing constant. The weight function $w(x)$ does not need to be bounded, $\int w(x)^2\tau(x)dx < \infty$ is sufficient if h is bounded. Still, if the normalized weights $w(Y^t)/\sum_s w(Y^s)$ are far from uniform, the above estimate becomes unreliable. We therefore need – like for accept/reject – a proposal that is not too far from the target.

If we want an unweighted sample instead of a weighted one, we can use resampling. We generate an additional sample (I^t) which takes values in $\{1, 2, \ldots, N\}$ with probabilities proportional to the weights $(w(Y^s))$:

$$\mathbb{P}(I^t = s) = \frac{w(Y^s)}{\sum_{r=1}^N w(Y^r)}$$

and set

$$Z^t = Y^{I_t}.$$

Thus values Y^t with large weights will be selected several times whereas those with small weights are likely to be not selected at all. This is easily seen to be correct because

$$\mathbb{E}\left(\sum_{t=1}^N h(Z^t) \mid Y^1, \ldots, Y^N\right) = \sum_{t=1}^N \sum_{s=1}^N \mathbb{P}(I^t = s) h(Y^s) = N \frac{\sum_{s=1}^N h(Y^s) w(Y^s)}{\sum_{r=1}^N w(Y^r)}.$$

Because the additional resampling step always increases the variance and it is not recommended to use it. However, in Section 4.4.4 we will see a situation where we need an unweighted sample to proceed in a recursive algorithm.

4.3 Basics of Markov chain Monte Carlo

As mentioned in the previous section, drawing an i.i.d sample from a complicated distribution π is difficult, especially in high dimensions. Markov chain Monte Carlo generates a sequence of random variables (X^t) which are dependent and such that the distribution of X^t converges weakly to π as $t \to \infty$. Estimation of $\int h(x)\pi(x)dx$ is still based on a law of large numbers, but now for dependent random variables:

$$\int h(x)\pi(x)dx \approx \bar{H}_{N,r} = \frac{1}{N-r} \sum_{t=r+1}^N h(X^t).$$

Here r is a "burn-in" period which discards values X^t whose distribution is too far from the target π.

The random variables are constructed recursively: The initial value X^0 is arbitrary, and for each $t \geq 1$ X^t is a deterministic function of X^{t-1} and a uniform random variable U^t which is independent of $X^0, \ldots, X^{t-1}$

$$X^t = G(X^{t-1}, U^t).$$

(In practice, often several uniform variables $U^{t,1}, U^{t,2}, \ldots, U^{t,k}$ are used, but this is equivalent). Because the dependence of X^t on previous random variables is only via X^{t-1}, the sequence (X^t) is called a Markov chain. The conditional distribution of X^t given X^{t-1} is called the transition kernel P of the chain

$$\mathbb{P}(X^t \in A \mid X^0, \ldots, X^{t-1}) = \mathbb{P}(X^t \in A \mid X^{t-1}) = P(X^{t-1}, A).$$

It is determined by the function G through

$$P(x, A) = \mathbb{P}(G(x, U) \in A) = \mathbb{P}(U \in \{u; G(x, u) \in A\}).$$

In particular, P does not depend on t because G is the same for all t. We therefore call the Markov chain time-homogeneous. In Markov process theory, one usually starts by specifying the transition kernel $P(x, A)$, the conditional probability that the next value of the chain is in A given that the current value is equal to x. It is always possible to construct a function G such that the above equation is satisfied. Because for Markov chain Monte Carlo, we need to draw from $P(x, .)$ for arbitrary values x, it is more natural to start with the concrete construction $X^t = G(X^{t-1}, U^t)$

In order to use Markov chain Monte Carlo to estimate expected values with respect to the target π, we need to find a transition kernel P such that we can draw from the distribution $P(x, .)$ for any x and such that for $N \to \infty$ the arithmetic mean of the $h(X^t)$ converges to $\int h(x)\pi(x)dx$. The general theory of Markov chains shows that the second requirement holds in a wide range of cases if the chain can reach all sets A with $\pi(A) > 0$ and if $X^{t-1} \sim \pi$ implies that $X^t \sim \pi$. If the second condition holds, we call π an invariant or stationary distribution for the transition kernel P. Because

$$\mathbb{P}(X^t \in A) = \mathbb{E}(\mathbb{P}(X^t \in A \mid X^{t-1})) = \mathbb{E}(\mathbb{P}(X^{t-1}, A)),$$

π is stationary for P if

$$\pi(A) = \int \pi(x)P(x, A)dx \quad \forall A,$$

or, in the case where $P(x, .)$ has the density $p(x, y)$, if

$$\pi(y) = \int \pi(x)p(x, y)dx.$$

There are two basic recipes for constructing a transition kernel P which has a given target distribution π as stationary distribution. The first one is the so-called Gibbs sampler. For this we assume that $x \in \mathbb{R}^p$ and we denote the conditional density of the i-th component of x, x_i given all the other components x, $x_{-i} = (x_j)_{j \neq i}$ by π_i:

$$\pi_i(x_i \mid x_{-i}) \propto \pi(x)$$

where $\propto$ means up to a term which does not contain x_i. This means that we can identify π_i by inspecting how the target density π depends on the i-th component. We don't need any integration. The densities π_i are also called "full conditionals" (because we condition on all other components). The Gibbs sampler depends on a "visiting schedule" $i_t \in \{1, 2, \ldots, p\}$ and iterates the following steps for $t = 1, 2, \ldots$

$$X^t_{i_t} \sim \pi_{i_t}(x_{i_t} \mid X^{t-1}_{-i_t})dx_{i_t}, \quad X^t_{-i_t} = X^{t-1}_{-i_t}.$$

In words, we leave all components of X^{t-1} except the one that is actually visited unchanged, and we update the visited component according to the conditional distribution of our target. By the definition of the conditional distribution, π is invariant for this transition kernel. The visiting schedule can be either deterministic or it can randomly select one of the components. In order that the chain can reach all sets, we have to visit each possible component infinitely often.

Example 4.2. *I.i.d. normal observations. Assume that the X_i are i.i.d $\sim \mathcal{N}(\mu, \frac{1}{\tau})$ and that we use the prior*

$$\pi(\mu, \tau) \propto \exp\left(-\frac{\kappa}{2}(\mu - \xi)^2\right) \tau^{\gamma-1} \exp(-\lambda\tau)$$

with hyperparameter $(\mu, \kappa, \gamma, \lambda)$. Because this is not a conjugate prior, the posterior

$$\pi(\mu, \tau \mid x_1, \ldots, x_n) \propto \tau^{n/2+\gamma-1} \exp\left(-\frac{\tau}{2}\sum(x_i - \mu)^2 - \lambda\tau - \frac{\kappa}{2}(\mu - \xi)^2\right)$$

is not a standard distribution. Because it is a two-dimensional distribution, we can get an idea about its shape by simply plotting the contours. The Gibbs sampler is also easy to use because

$$\begin{aligned}\pi(\mu \mid \tau, x_1, \ldots, x_n) &\propto \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\kappa}{2}(\mu-\xi)^2\right) \propto \exp\left(-\frac{n\tau}{2}(\mu-\bar{x})^2 - \frac{\kappa}{2}(\mu-\xi)^2\right)\\ &\propto \exp\left(-\frac{n\tau+\kappa}{2}\left(\mu - \frac{n\tau}{n\tau+\kappa}\bar{x} - \frac{\kappa}{n\tau+\kappa}\mu_0\right)^2\right)\end{aligned}$$

and

$$\pi(\tau \mid \mu, x_1, \ldots, x_n) \propto \tau^{n/2+\alpha-1}\exp\left(-\tau\left(\lambda + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right).$$

Therefore the Gibbs sampler alternates between drawing μ from a normal distribution and τ from a Gamma distribution. We will illustrate this by a computer demonstration in the lecture.

The Gibbs sampler requires that we can sample from the full conditionals. Because these distributions are one-dimensional, this is often possible. If it isn't, we can use the Metropolis-Hastings algorithm instead. It is based on the fact that any reversible distribution is also stationary. Here π is called reversible for the transition kernel P if

$$\int_A \pi(x)P(x,B)dx = \int_B \pi(x)P(x,A)dx \quad \forall A, B,$$

or in other words, if $X^t \sim \pi$, then

$$\mathbb{P}(X^t \in A, X^{t+1} \in B) = \mathbb{P}(X^{t+1} \in A, X^t \in B) \quad \forall A, B.$$

Choosing for B the whole space $\mathbb{R}^p$, it follows that a reversible dsitribution π is stationary. If $P(x,.)$ has the density $p(x,y)$ for any x, then reversibility is equivalent to

$$\pi(x)p(x,y) = \pi(y)p(y,x) \quad \forall x, y.$$

For any pair $x \neq y$, we can therefore choose one of the two values $p(x,y)$ and $p(y,x)$ arbitrarily, whereas the other one is determined by the reversibility equation. However, a solution obtained in this way does in general not satisfy $\int p(x,y)dy = 1$ for any x and thus is not the density of a transition kernel. To solve this problem, one can start with an arbitrary transition density q and then choose from the two possible solutions

$$p(x,y) = q(x,y),\ p(y,x) = \frac{\pi(x)q(x,y)}{\pi(y)}$$

and

$$p(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)},\ p(y,x) = q(y,x)$$

the one which satisfies $p(x,y) \leq q(x,y)$ and $p(y,x) \leq q(y,x)$ for any $x \neq y$. This solution can be written in the compact form

$$p(x,y) = q(x,y)\min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right).$$

It folows that $\int p(x,y)dy \leq \int q(x,y)dy = 1$ for any x, and one can put the missing mass on the diagonal, meaning that the chain does not move. Written in formulae, the transition kernel is

$$P(x,A) = \int_A p(x,y)dy + 1_A(x)\left(1 - \int p(x,y)dy\right).$$

Assuming that we can simulate from the transition density $q(x,.)$ for any x, the following algorithm generates a Markov chain with the transition kernel P:

- At time t generate $Y^t \sim q(X^{t-1},x)dx$ and $U^t \sim$ uniform(0,1), independently from each other and independently of previously generated variables.
- Set
$$X^t = \begin{cases} Y^t & \text{if} \quad U^t \leq \min\left(1, \frac{\pi(Y^t)q(Y^t,X^{t-1})}{\pi(X^t)q(X^{t-1},Y^t)}\right) \\ X^{t-1} & \text{else} \end{cases}$$

This is similar to the accept/reject method, but the proposal depends on the most recent value, and in case of a rejection, we do not move. The simplest choice of $q(x,.)$ is a normal density with mean x and an arbitrary covariance matrix Σ. In this case, $q(x,x') = q(x',x)$ so that the acceptance probability is simply $\min(1, \pi(x')/\pi(x))$. This means moving to value which is more likely than the current value is always accepted whereas the acceptance of a move to a less likely value is given by the likelihood ratio. The algorithm for a symmetric q is due to Metropolis et al. (1953) whereas the general case is due to Hastings (1970).

One can also combine Metropolis-Hastings with Gibbs by proposing a move in only one or a few components of x. The formula for accepting a move has the same structure as before.

The Gibbs and Metropolis algorithms are very flexible and can in principle handle most problems. However, judging how reliable the results are is not always easy. The situation differs from independent Monte Carlo in two aspects. First, because X^t is only approximately distributed according to the target π, there is a bias:

$$\mathbb{E}(\bar{H}_{N,r}) \neq \int h(x)\pi(x)dx.$$

Second, because successive values X^t are dependent, the variance is more complicated

$$\text{Var}\left(\bar{H}_{N,r}\right) = \frac{1}{(N-r)^2}\left(\sum_{t=r+1}^{N} \text{Var}(h(X^t)) + 2\sum_{t=r+1}^{N}\sum_{s=1}^{N-t} \text{Cov}(h(X^t), h(X^{t+s}))\right).$$

The pragmatic way to deal with these complications is to look at the times series plots of $h(X^t)$ or of components X_i^t versus t and to choose r such that the series "looks stationary" for $t \geq r$. Then one assumes that $X^t \sim \pi$ for $t > r$ so that there is no bias, $\text{Var}(h(X^t))$ is independent of t and the covariances $\text{Cov}(h(X^t), h(X^{t+s}))$ depend only on s and can be estimated by

$$\frac{1}{N-r}\sum_{t=r+1}^{N-s}(h(X^t) - \bar{H}_{r,N})(h(X^{t+s}) - \bar{H}_{N,r}).$$

The number of replicates N should then be large enough that these estimated covariances are close to zero for most lags s.

4.4 Some advanced computational methods

4.4.1 Adaptive MCMC

We have seen in the previous section that usually there are many transition kernels which have a given target π as the stationary distribution. The prime example is the Metropolis-Hastings algorithm with proposal $Y^t \sim \mathcal{N}(X^{t-1}, \Sigma)$ where Σ is an arbitrary positive definite covariance matrix. Although any choice of Σ leads to a consistent estimation of the quantities of interest $\int h(x)\pi(x)dx$, the choice of Σ has a large influence on the quality of the approximation for any finite number N of steps. In some prototype cases, it has been shown that if π is a p-dimensional distribution with covariance matrix $\mathrm{Cov}_\pi(X)$, then the "optimal" choice of Σ is given by

$$\Sigma = \frac{2.38^2}{p} \mathrm{Cov}_\pi(X).$$

A similar result says that the "optimal" choice is such that the average acceptance rate – after we have reached the stationary distribution – is 0.234, that is

$$\int \pi(x) \int q(x,y) \min\left(1, \frac{\pi(y)}{\pi(x)}\right) dydx = 0.234.$$

These criteria are considered as reasonable rules of thumb also in cases which differ from the ones in which they have been derived. The problem is however that they cannot be used directly because they depend on the unknown target π. A standard strategy is to have first an exploration phase where one tries out various values of Σ. In a second phase one then runs the algorithm with a fixed Σ determined from the experience gained in the exploration phase.

Adaptive MCMC combines the two phases by using a time dependent Σ^t which depends on the sequence of values $(X^0, X^1, \ldots, X^{t-1})$ generated so far. For instance, in order to approximate the first criterion mentioned above, one can take

$$\Sigma^t = \frac{2.38^2}{p} \frac{1}{t-1} \sum_{s=0}^{t-1} (X^s - \bar{X}^{t-1})(X^s - \bar{X}^{t-1})^T, \quad \bar{X}^{t-1} = \frac{1}{t} \sum_{s=0}^{t-1} X^s.$$

For the second criterion, let us assume that we only want to optimise the scale of Σ where the shape is fixed, e.g. $\Sigma = \sigma^2 I_p$. If we assume that the acceptance probability is a decreasing function of σ^2, then the following rule to choose $\sigma^{2,t}$ seems reasonable

$$\sigma^{2,t} = \begin{cases} r_t \sigma^{2,t-1} & \text{if} \quad \frac{1}{t-1} \sum_{s=0}^{t-2} \min(1, \frac{\pi(Y^{s+1})}{\pi(X^s)}) > 0.234, \\ \frac{1}{r_t} \sigma^{2,t-1} & \text{if} \quad \frac{1}{t-1} \sum_{s=0}^{t-2} \min(1, \frac{\pi(Y^{s+1})}{\pi(X^s)}) < 0.234. \end{cases}$$

Here Y^s is the proposed value in step s and $r_t \downarrow 1$.

The theoretical analysis of these and similar algorithms is delicate because one has to control the probability that this algorithm results in choosing a sequence Σ^t where some eigenvalues go to zero or to infinity. For more details, discussion and examples, see C. Andrieu, J. Thoms, A tutorial on adaptive MCMC, Stat. Comput. (2008) 18, 343 - 373.

4.4.2 Hamiltonian Monte Carlo

This is a method which allows the chain to make big moves that are still accepted with high probability. The price to pay is that we have to be able to evaluate the gradient of $\log \pi$ efficiently. The method is based on the observation that a deterministic transition $X^t = G(X^{t-1})$ has π as stationary distribution if it is volume preserving and keeps π invariant:

$$\pi(G(x)) = \pi(x), \quad \left| \det \frac{\partial G(x)}{\partial x} \right| = 1 \quad \forall x.$$

This follows immediately from the transformation formula for densities. With this transition we cannot reach all possible sets A with $\pi(A) > 0$ and thus it does not satisfy our second condition. Moreover, it seems difficult to find a map G with these two properties. However, both problems can be avoided if we consider a new target distribution on a space with doubled dimension:

$$\tilde{\pi}(x,u) \propto \pi(x) \exp\left(-\sum_{i=1}^{p} \frac{u_i^2}{2m_i} \right).$$

Clearly, if $(X,U) \sim \tilde{\pi}$, then $X \sim \pi$ and if morever $U' \sim \mathcal{N}(0, \mathrm{diag}(m_i))$ is independent of (X,U), then also $(X,U') \sim \tilde{\pi}$. Hence if we alternate between drawing an independent new component u and applying a transformation G which is volume preserving and keeps $\tilde{\pi}$ invariant, we obtain a Markov chain which satisfies both conditions. Moreover, both transitions can make big moves in the (x,u) space and thus the chain converges quickly to the stationary distribution and has autocorrelations which decay quickly.

The construction of the map which is volume preserving and keeps $\tilde{\pi}$ invariant is based on Hamiltonian mechanics which led to the name “Hamiltonian Monte Carlo”. We define the Hamiltonian

$$H(x,u) = -\log \pi(x) + \sum_{i=1}^{p} \frac{u_i^2}{m_i}.$$

In physical terms, x is the position, u is the momentum, $-\log \pi(x)$ is the potential and $\sum \frac{u_i^2}{m_i}$ the kinetic energy. The transformation $G(x,u)$ is then the solution at time T of the ordinary differential equation

$$\frac{dx_i}{dt} = \frac{\partial H(x,u)}{\partial u_i}, \quad \frac{du_i}{dt} = -\frac{\partial H(x,u)}{\partial x_i}$$

with initial condition (x,u). The value of T and of the “masses” m_i are parameters of the method, to be chosen by the user. It is well known from basic mathematical physics (and straightforward to verify) that

$$\frac{d}{dt} H(x(t), u(t)) = 0, \quad \nabla \left(\frac{\partial H(x,u)}{\partial u}, -\frac{\partial H(x,u)}{\partial x} \right) = 0.$$

The first equation shows that G leaves $\tilde{\pi}$ invariant and the second that G is volume preserving. In addition, the inverse transformation G^{-1} is obtained by changing the sign on the right hand side of the differential equation. By the form of the Hamiltonian, it follows that $G^{-1}(x,u) = (G(x,-u)_x, -G(x,-u)_u)$, that is up to a reflection of the momentum, the time evolution is reversible.

For implementation, we need to solve the above differential equation by some discretization procedure. This unfortunately has the consequence that the transformation does not keep

$\tilde{\pi}$ exactly invariant. The so-called "leap frog method" however induces only small changes to $\tilde{\pi}$, it preserves volume exactly and it is time reversible (up to reflecting u). Thus it is possible to restore the exact invariance of $\tilde{\pi}$ by a Metropolis-Hastings acceptance step at the end.

In order to see the advantage of the Hamiltonian Monte Carlo, let us assume that at some time t $u_i > 0$, that is the i-th component of x moves to the right. If $\partial(\log \pi(x))/\partial x_i > 0$ the i-th component moves into a region which is more likely under the target π, and by the Hamiltonian dynamics the momentum u_i increases and so we move faster. On the other hand, if $\partial(\log \pi(x))/\partial x_i < 0$ we are going into a less likely region and the momentum u_i decreases and eventually gets reversed.

For more details, see e.g. Chapter 5 in Handbook of Markov Chain Monte Carlo, S. Brooks, A. Gelman, G.L. Jones and X.L. Meng (eds), Chapman and Hall, 2011.

4.4.3 Reversible jump MCMC

So far our target distribution π always was a distribution on $\mathbb{R}^p$ for some fixed dimension p. In Bayesian model selection or nonparametric function estimation we are interested in a model indicator $k = 1, 2, \ldots$ and a parameter vector $\theta_k \in \mathbb{R}^{p_k}$. In nonparametric function estimation, k is typically the number of basis elements in an expansion of the unknown function with respect to a given basis. For instance, if we use splines k would be the number of knots and θ_k would contain the location of the knots and the coefficients of the function in the B-spline basis. Or if we use Gaussian mixtures to approximate arbitrary densities, k is the number of components in the mixture and θ_k contains the means and covariances of the k components. In such situations the posterior $\pi(\theta_k, k \mid x)$ is thus a distribution on the space

$$\mathbb{X} = \cup_k \mathbb{R}^{p_k}$$

where the union may be over a finite or a countable set. In the following we denote elements of $\mathbb{X}$ by (x_k, k) and we consider an arbitrary target distribution π on $\mathbb{X}$ which need not be a posterior distribution coming from a Bayesian analysis.

The first idea is to use a separate MCMC algorithm on $\mathbb{R}^{p_k}$ to draw from $\pi(x_k \mid k) \propto \pi(x_k, k)$ for each k. However, this does not give us information about the probabilities

$$\pi(k) = \int_{R^{p_k}} \pi(x_k, k) dx_k$$

for the model indicator k, and it is computationally inefficient to spend a lot of effort for those k where $\pi(k)$ is small. Reversible jump MCMC draws directly from the full distribution of both k and x_k so that $\pi(k)$ is approximated by the number of draws that lie in $\mathbb{R}^{p_k}$. For this we need an algorithm which proposes also jumps from $\mathbb{R}^{p_k}$ to some $\mathbb{R}^{p_m}$ for $m \neq k$ in order to explore the whole state space. We want to use the basic Metropolis-Hastings idea where we propose a value based on some arbitrary proposal distribution and then use an accept/reject step to ensure that the target π is stationary. A closer analysis shows that if the proposal allows jumping from $\mathbb{R}^{p_k}$ to $\mathbb{R}^{p_m}$, then the reverse jump from $\mathbb{R}^{p_m}$ to $\mathbb{R}^{p_k}$ must also be allowed and we must have the same degrees of freedom for the pairs (x_k, x_m) and (x_m, x_k). More precisely, assume that from a given $x_k \in \mathbb{R}^{p_k}$ we propose

$$X_m = x_m(x_k, U_{mk}), \quad U_{mk} \in R^{d_{mk}} \sim f_{mk}(u)du$$

and from a given $x_m \in \mathbb{R}^{p_m}$ we propose

$$X_k = x_k(x_m, U_{km}), \quad U_{km} \in R^{d_{km}} \sim f_{km}(u)du.$$

Then the dimensions must match in the sense that

$$p_k + d_{mk} = p_m + d_{km}$$

and there must be a bijection between (x_k, u_{mk}) and (x_m, u_{km}). The acceptance probability contains then also the Jacobi determinant

$$|\det\left(\frac{\partial(x_m, u_{km})}{\partial(x_k, u_{mk})}\right)|.$$

For more details, see e.g. Chapter 6 in Highly Structured Stochastic Systems, P.J. Green, N.L. Hjort and S. Richardson (eds), Oxford University Press, 2003.

4.4.4 Sequential Monte Carlo

This is a name used for methods which sample not from one target π, but from a sequence of related targets $\pi_0, \pi_1, \dots \pi_n$. For instance, we can take as π_k the posterior of θ given the first k observations, or we can take

$$\pi_k(x)dx \propto \pi(x)^{\phi_k}, \quad \phi_0 < \phi_1 < \dots \phi_n = 1.$$

This means that π_0 is close to a uniform distribution and we can for instance use accept/reject to generate a sample $(X^{0,t})$ from π_0. This sample is then sequentially modified so that at the end we have a sample $(X^{n,t})$ from the original target π. In particular, if π is multimodal, then this method is preferable over the Metropolis algorithm because it has better chances to sample from regions around all modes.

The modification occurs by a propagation and a reweighting/resampling step. Propagation means that at stage k we choose a transition density p_k and generate

$$Y^{k,t} \sim p_k(X^{k-1,t}, y)dy, \quad \text{independently for } t = 1, 2, \dots N.$$

Then

$$Y^{k,t} \sim \int \pi_{k-1}(x)p_k(x,y)dx \cdot dy,$$

so in order to transform $(Y^{k,t})$ into a sample from π_k by importance sampling, we have to use the weights

$$w^{k,t} \propto \frac{\pi_k(Y^{k,t})}{\int \pi_{k-1}(x)p_k(x, Y^{k,t})dx}.$$

However, the integral in the denominator is in general not available analytically and thus the method cannot be used. The key idea is to look at the pairs $(X^{k-1,t}, Y^{k.t})$ whose density is equal to $\pi_{k-1}(x)p_k(x,y)$. We want to convert this distribution into one which has $\pi_k(y)$ as the second marginal. But all such densities have the form $\pi_k(y)q_{k-1}(y,x)$ where q_{k-1} is an arbitrary transition density. Hence if we set

$$X^{k,t} = Y^{k,I^t}, \quad \mathbb{P}(I^t = s) \propto \frac{\pi_k(Y^{k,s})q_{k-1}(Y^{k,s}, X^{k-1,s})}{\pi_{k-1}(X^{k-1,t})p_k(X^{k-1,t}, Y^{k,t})},$$

$(X^{k,t})$ is a sample from π_k.

If one is willing to use weighted samples at all stages, then we do not need a resampling step: We would then simply update the weights $w^{k,t}$

$$w^{k,t} \propto w^{k-1,t}\frac{\pi_k(Y^{k,s})q_{k-1}(Y^{k,s}, X^{k-1,s})}{\pi_{k-1}(X^{k-1,t})p_k(X^{k-1,t}, Y^{k,t})}.$$

However, this sequential multiplication leads very quickly to unbalanced weights, and adding the resampling step gives more reliable approximations: The ties which arise from the resampling at step k disappear in the next propagation step. Resampling helps to concentrate the computing effort in those region of the space where the densities π_k have their main mass.

For more details see e.g. P. Del Moral, A. Doucet and A. Jasra, Sequential Monte Carlo Samplers, J. Royal Statist. Soc. B 68 (2006): 411-436.

4.4.5 Approximate Bayesian computation

For some models, evaluating the likelihood $f(x \mid \theta)$ is complicated or even impossible. In such cases, neither the Gibbs sampler nor Metropolis Hastings can be used to simulate from the posterior

$$\pi(\theta \mid x_{obs}) \propto \pi(\theta) f(x_{obs} \mid \theta).$$

But often, simulating a random variable $X \sim f(x \mid \theta)dx$ is much easier and we can therefore generate pairs $(\theta^t, X^t) \sim \pi(\theta) f(x \mid \theta) d\theta dx$. If X is discrete, we can use the accept/reject method to simulate from the density proportional to

$$\pi(\theta) f(x \mid \theta) 1_{[x=x_{obs}]}$$

whose marginal is the posterior: We simply accept only pairs (θ^t, X^t) such that $X^t = x_{obs}$. Of course, it might take a very long time until we accept any pair at all, and in most cases X is continuous anyhow. But we can use the same idea if we replace the point mass at x_{obs} by a distribution which is concentrated near x_{obs}

$$\pi(\theta) f(x \mid \theta) \exp(-d(x, x_{obs})/\varepsilon)$$

where d is a metric on the space of observations. Instead of working with a fixed ϵ, one can also choose a sequence $\varepsilon_n \to 0$ with a rather large ε_0 and use a sequential Monte Carlo algorithm to produce samples of the corresponding targets.

For more details see e.g. M. Marin, P. Pudlo, C.P. Robert, and R.J. Ryder, Approximate Bayesian computational methods, Statistics and Computing, 22 (2012): 1167 - 1180.