

MATH 559

Bayesian Theory and Methods

Dr David A. Stephens

Department of Mathematics & Statistics
Room 1225, Burnside Hall
david.stephens@mcgill.ca

Part 1

Bayesian Theory

Probability Fundamentals

Introduction

The objective of a statistical analysis is to use data to make *optimal* and *coherent* decisions, including

- *inference*: making statements about the unknown data generating mechanism;
- *prediction*: making statements about as yet unobserved ('future') data,

whilst appropriately representing the uncertainty associated with these decisions.

Typically, the analysis is based on a probabilistic (or statistical) model.

Notation and Basic Concepts

Let Y denote a single random variable taking values on $\mathcal{Y} \subseteq \mathbb{R}$.

- Y records the result of some measurement procedure;
- \mathcal{Y} could be countable (so that Y is ‘discrete’).

Let y denote an observed value associated with Y .

Notation and Basic Concepts

A *probability model* for Y is encapsulated in a *probability function*, $P_Y(\cdot)$, where (informally) for set $A \subseteq \mathbb{R}$,

$$P_Y(A) \equiv \Pr[Y \in A]$$

and more specifically

$$P_Y ((-\infty, c]) \equiv \Pr[Y \leq c].$$

We define the *distribution function*, $F_Y(\cdot)$, via the specification

$$F_Y(c) \equiv P_Y ((-\infty, c]) = \Pr[Y \leq c] \quad c \in \mathbb{R}.$$

Notation and Basic Concepts

If Y is *discrete*, then

$$\mathcal{Y} = \{y_1^*, y_2^*, \dots, \}$$

and for any $y \in \mathbb{R}$, we have

$$F_Y(y) = \sum_{j: y_j^* \leq y} \Pr[Y = y_j^*] = \sum_{j: y_j^* \leq y} p_Y(y_j^*)$$

say, where

$$p_Y(y_j^*) = \Pr[Y = y_j^*]$$

is the *probability mass function* (pmf) for Y .

Notation and Basic Concepts

If we can write

$$F_Y(y) = \int_{-\infty}^y f_Y(t) \, dt \quad y \in \mathbb{R}$$

then we term F_Y an *absolutely continuous distribution*, with

$$f_Y(y)$$

the *probability density function* (pdf) for Y . In this case

$$f_Y(y) = \left. \frac{dF_Y(t)}{dt} \right|_{t=y}$$

Notation and Basic Concepts

For simplicity we can unify notation for the discrete and continuous cases by writing

$$\Pr[Y \in A] = \int_A F_Y(dy) \equiv \begin{cases} \sum_{y \in A} p_Y(y) & Y \text{ discrete} \\ \int_A f_Y(y) dy & Y \text{ continuous} \end{cases}$$

Notation and Basic Concepts

$F_Y(y)$ can have both discrete and continuous components.

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{2} & y = 0 \\ \frac{1}{2} + \frac{1}{2}(1 - e^{-y}) & y > 0 \end{cases}$$

Notation and Basic Concepts

We have

$$\Pr[Y = 0] = \Pr[Y > 0] = \frac{1}{2}$$

and for $y > 0$

$$\begin{aligned}\Pr[Y \leq y | Y > 0] &= \frac{\Pr[Y \leq y, Y > 0]}{\Pr[Y > 0]} = \frac{\Pr[0 < Y \leq y]}{\Pr[Y > 0]} \\ &= \frac{\Pr[Y \leq y] - \Pr[Y \leq 0]}{\Pr[Y > 0]} \\ &= 1 - e^{-y}\end{aligned}$$

Notation and Basic Concepts

The notations

$$\Pr[Y \in A] = \int_A F_Y(dy) = \int_A dF_Y(y)$$

are also sometimes used. Informally

$$dF_Y(y) = \lim_{h \rightarrow 0^+} (F_Y(y+h) - F_Y(y))$$

Notation and Basic Concepts

In practice, we observe data ('*observables*')

$$y_1, \dots, y_n$$

and use them to learn about the unknown ('*unobservable*') model P_Y or F_Y , or features of it such as its expectation

$$\theta = \int y F_Y(dy)$$

That is, it is the distribution F_Y that is unknown.

Notation and Basic Concepts

The data are realizations of random variables Y_1, \dots, Y_n , and we have observed the event

$$\Pr \left[\bigcap_{i=1}^n (Y_i \in \{y_i\}) \right].$$

This is a *joint* probability, so we need to consider the *joint probability model*

$$\Pr \left[\bigcap_{i=1}^n (Y_i \in A_i) \right]$$

for arbitrary subsets A_1, \dots, A_n of \mathbb{R} .

Notation and Basic Concepts

Specifically, we consider the *joint cdf*

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] \quad (y_1, \dots, y_n) \in \mathbb{R}^n$$

or quantities derived from it (*joint pdf* etc).

Notation and Basic Concepts

The laws of probability dictate that joint distributions must obey certain *coherence* properties, such as *marginalization* results: for example

$$F_{Y_1, Y_2}(y_1, y_2) = \lim_{y_3 \rightarrow \infty} F_{Y_1, Y_2, Y_3}(y_1, y_2, y_3)$$

i.e. the joint distribution of (Y_1, Y_2, Y_3) *automatically* specifies the joint distributions of all subsets of the variables.

Notation and Basic Concepts

A typical assumption in statistical settings is that Y_1, \dots, Y_n are *independent*, that is that for all $(y_1, \dots, y_n) \in \mathbb{R}^n$

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \prod_{i=1}^n \Pr [Y_i \leq y_i]$$

so that

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \prod_{i=1}^n F_{Y_i}(y_i).$$

Notation and Basic Concepts

Further, it is often assumed that the Y_1, \dots, Y_n are *identically distributed*

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \prod_{i=1}^n F_Y(y_i).$$

However, these are quite strong assumptions.

Notation and Basic Concepts

A weaker assumption is that of *infinite exchangeability*: we consider an infinite sequence

$$Y_1, Y_2, Y_3, \dots$$

for which, *for all* $n \geq 1$ and sets A_1, \dots, A_n we have that

$$\Pr \left[\bigcap_{i=1}^n (Y_i \in A_i) \right] = \Pr \left[\bigcap_{i=1}^n (Y_i \in A_{\sigma(i)}) \right]$$

for all permutations $(\sigma_{(1)}, \dots, \sigma_{(n)})$ of indices $(1, \dots, n)$.

Notation and Basic Concepts

$n = 2$:

$$\Pr[(Y_1 \leq y_1) \cap (Y_2 \leq y_2)] = \Pr[(Y_1 \leq y_2) \cap (Y_2 \leq y_1)]$$

$n = 3$:

$$\begin{aligned} &\Pr[(Y_1 \leq y_1) \cap (Y_2 \leq y_2) \cap (Y_3 \leq y_3)] \\ &= \Pr[(Y_1 \leq y_2) \cap (Y_2 \leq y_1) \cap (Y_3 \leq y_3)] \\ &= \Pr[(Y_1 \leq y_3) \cap (Y_2 \leq y_2) \cap (Y_3 \leq y_1)] \\ &= \Pr[(Y_1 \leq y_3) \cap (Y_2 \leq y_1) \cap (Y_3 \leq y_2)] \\ &= \dots \end{aligned}$$

Notation and Basic Concepts

For infinite exchangeability: need the permutation equality

- (a) to hold for any *finite* n drawn from the infinite sequence
i.e. for any single infinite collection, we need the equality of the probabilities to hold for any finite subset of the collection.

Notation and Basic Concepts

(b) to respect *marginalization* conditions; that is

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right]$$

must be compatible with

$$\Pr \left[\bigcap_{i=1}^{n+1} (Y_i \leq y_i) \right]$$

in the sense that

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \lim_{y_{n+1} \rightarrow \infty} \Pr \left[\bigcap_{i=1}^{n+1} (Y_i \leq y_i) \right]$$

Notation and Basic Concepts

Therefore it is not straightforward to build joint distributions for exchangeable sequences of random variables.

- need to make the distribution for $n = 1$ compatible with the distribution for $n = 2, 3, \dots$;
- this is more complicated than assuming simple independence and identical distribution.

Notation and Basic Concepts

Example: Binary case

Suppose we have an infinitely exchangeable sequence $\{Y_n\}$, where for each i , $Y_i \in \{0, 1\}$. Consider for $n \geq 1$

$$\Pr[(Y_1 = y_1) \cap \cdots \cap (Y_n = y_n)]$$

which we may write in short

$$\Pr[Y_1 = y_1, \dots, Y_n = y_n],$$

where we consider vector arguments

$$(y_1, \dots, y_n) \in \{0, 1\}^n$$

Notation and Basic Concepts

Example: Binary case

Then under infinite exchangeability, we must have that

$$\Pr[Y_1 = y_1, \dots, Y_n = y_n]$$

depends only on the value of

$$s_n = \sum_{i=1}^n y_i.$$

There are 2^n possible binary vectors of length n , but

$$s_n \in \{0, 1, \dots, n\}$$

so there are a maximum of $(n + 1)$ different probabilities, although these probabilities must sum to 1.

Notation and Basic Concepts

Example: Binary case

- $n = 1$: $s_n \in \{0, 1\}$, so denote the probabilities $p_{1,0}$ and $p_{1,1}$, where we must have

$$p_{1,0} = 1 - p_{1,1}$$

Notation and Basic Concepts

Example: Binary case

- $n = 2$: $s_n \in \{0, 1, 2\}$, so denote the probabilities $p_{2,0}$, $p_{2,1}$ and $p_{2,2}$, where we must have that

$$p_{2,0} = 1 - p_{2,1} - p_{2,2}$$

but also due to marginalization that

$$\begin{aligned} p_{1,y_1} &= \Pr[Y_1 = y_1] \\ &= \Pr[Y_1 = y_1, Y_2 = 0] + \Pr[Y_1 = y_1, Y_2 = 1] \\ &= p_{2,y_1} + p_{2,y_1+1} \end{aligned}$$

for $y_1 \in \{0, 1\}$.

Notation and Basic Concepts

Example: Binary case

This construction be extended to define the required relations for any n .

However, to specify the distribution in this way, we need to specify and compute the relations *for all* n .

Notation and Basic Concepts

Example: Binary case

Assuming *independence*, we have

$$\Pr[Y_1 = y_1, \dots, Y_n = y_n] = \prod_{i=1}^n \Pr[Y_i = y_i] = p^{s_n} (1 - p)^{n - s_n}$$

where

$$p = \Pr[Y_i = y_i] \quad i = 1, \dots, n.$$

Conditional probability

For two events, E_1, E_2 with $P(E_2) > 0$, we have that

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

is the *conditional probability* for E_1 given E_2 .

- $P(E_1)$ is the probability that E_1 occurs;
- $P(E_1|E_2)$ is the probability that E_1 occurs *if we have information* that E_2 occurs.

Conditional probability

- *relative* to the probability of E_2 , what is the probability that *both* E_1 and E_2 occur ?
- E_1 and E_2 are *independent* if and only if

$$P(E_1|E_2) = P(E_1)$$

Conditional probability

For two events, E_1, E_2 with both $P(E_1) > 0$ and $P(E_2) > 0$, we have by the definition that

$$P(E_1|E_2) = \frac{P(E_2|E_1)P(E_1)}{P(E_2)}.$$

We know this result as *Bayes Theorem*.

Conditional probability

Exchangeability assumptions allow for *dependence*: that is, for example

$$\Pr[Y_{n+1} \in A_{n+1} | Y_1 = y_1, \dots, Y_n = y_n]$$

does not reduce to

$$\Pr[Y_{n+1} \in A_{n+1}]$$

as in the independence case. That is, for all i and j , Y_i and Y_j are identically distributed, but not independent.

Conditional probability

Example: Infinitely exchangeable binary case

$$\begin{aligned}\Pr[Y_{n+1} = 1 | Y_1 = y_1, \dots, Y_n = y_n] \\&= \frac{\Pr[Y_1 = y_1, \dots, Y_n = y_n, Y_{n+1} = 1]}{\Pr[Y_1 = y_1, \dots, Y_n = y_n]} \\&= \frac{p_{n+1, s_n+1}}{p_{n, s_n}}\end{aligned}$$

where

$$s_n = \sum_{i=1}^n y_i.$$

Conditional probability

Example: Independent binary case

$$\begin{aligned}\Pr[Y_{n+1} = 1 | Y_1 = y_1, \dots, Y_n = y_n] \\&= \frac{\Pr[Y_1 = y_1, \dots, Y_n = y_n, Y_{n+1} = 1]}{\Pr[Y_1 = y_1, \dots, Y_n = y_n]} \\&= \frac{p^{s_n+1}(1-p)^{n-s_n}}{p^{s_n}(1-p)^{n-s_n}} \\&= p \\&= \Pr[Y_{n+1} = 1]\end{aligned}$$

Conditional probability

Note

It is possible to consider *finite exchangeability*, where the exchangeability holds for a *finite* collection of random variables, that is, for a *specific* $n \geq 1$

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \Pr \left[\bigcap_{i=1}^n (Y_i \leq y_{\sigma(i)}) \right]$$

for all permutations $(\sigma_{(1)}, \dots, \sigma_{(n)})$ of indices $(1, \dots, n)$.

Inference and Prediction

Inference and prediction

In *statistical* calculations

- we observe data y_1, \dots, y_n and wish to make statements about unknown quantities in light of the data;
- given the data, *what do we think about the model* ?

If F_Y is *known*, there is no *inference* problem, and prediction can be carried out via F_Y .

Inference and prediction

If F_Y is *unknown*, then it is the focus of our inference.

- we treat F_Y as an unknown, and make statements about it in light of the data;
- given the data, *what do we think about F_Y ?*

We treat F_Y as a *random variable*.

Inference and prediction

If F_Y is *unknown*, then an independent and identically distributed (IID) statement of the sort above is really a *conditional* statement given F_Y .

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \middle| F_Y \right] = \prod_{i=1}^n F_Y(y_i)$$

Inference and prediction

If F_Y is a random variable, we must be able to specify a probability distribution for it:

- in general, F_Y is an *infinite-dimensional* object;
- F_Y has certain specific properties.

Need the capability to build a probability distribution on the space of functions. \mathcal{F} say, that satisfy the properties of distribution functions.

Parametric modelling

The most common approach involves using a *finite* dimensional '*parameter*', $\theta \in \mathbb{R}^d$ say, and specifying that

$$F_Y(y) \equiv F_Y(y; \theta) \quad y \in \mathbb{R}$$

so that the unknown quantity is now θ , and $F_Y(\cdot; \theta)$ is a *known functional form*. Then

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \middle| \theta \right] = \prod_{i=1}^n F_Y(y_i; \theta)$$

Non-parametric modelling

The *non-parametric* approach involves using an *infinite* dimensional parameter, the function $F_Y(\cdot)$ itself. We write

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \middle| F_Y \right] = \prod_{i=1}^n F_Y(y_i)$$

Semi-parametric modelling

The *semi-parametric* approach involves using a model that is specified in terms of both

- a *finite* dimensional parameter, $\theta \in \mathbb{R}^d$
- an *infinite* dimensional parameter.

Semi-parametric modelling

Example: Semi-parametric location model

The model

$$F_Y(y; \theta) \equiv F(y - \theta)$$

where $\theta \in \mathbb{R}$ and F is an arbitrary cdf is a semi-parametric location model for a univariate random variable Y .

θ locates the distribution on the y axis, but we may want to leave the 'shape' of the distribution as general as possible.

Interest and nuisance parameters

We often partition parameters into two components

- *parameters of interest*: the focus of inference;
- *nuisance parameters*: parameters necessary for the specification of the probability model, but which are not the focus of interest.

Interest and nuisance parameters

In the parametric case, we might partition $\theta = (\psi, \lambda)$ where ψ is the parameter of interest.

In the semi-parametric case, the non-parametric component is often regarded as a nuisance parameter.

Interest and nuisance parameters

In the non-parametric case, inference focusses on F_Y itself, or possibly some ‘functional’ of F_Y .

For example, we might target the *expectation* of F_Y

$$\mu(F_Y) = \mathbb{E}_Y[Y; F_Y] = \int y \, dF_Y(y) \equiv \int y \, F_Y(dy)$$

if this quantity is finite.

de Finetti's Representation

de Finetti's Representation

A key result of Bayesian theory defines the probability distribution of infinitely exchangeable random variables:

- this characterizes all possible forms for the distribution;
- it gives a straightforward mechanism for the construction of arbitrary distributions for infinitely exchangeable sequences;
- this result underpins the logic of Bayesian inference and prediction.

de Finetti's Representation

We consider this result in two forms

- when the observables are all binary, and
- when there is no explicit assumption about the observables.

de Finetti's Representation

Theorem: 0-1 representation theorem

Suppose that Y_1, Y_2, \dots is an infinitely exchangeable sequence of 0-1 variables. Then there exists a distribution function $\pi_0(\cdot)$ such that for all $n \geq 1$, the joint mass function of (Y_1, Y_2, \dots, Y_n) can be represented

$$p_{Y_1, Y_2, \dots, Y_n}(y_1, y_2, \dots, y_n) = \int_0^1 \left\{ \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \pi_0(d\theta)$$

for some probability distribution $\pi_0(\cdot)$.

de Finetti, Hewitt-Savage

de Finetti's Representation

Theorem: 0-1 representation theorem

Furthermore, $\pi_0(\cdot)$ is defined for $0 \leq \theta \leq 1$ by

$$\int_0^\theta \pi_0(dt) = \lim_{n \rightarrow \infty} \Pr[R_n \leq \theta] \quad (\diamond)$$

and where

$$S_n = \sum_{i=1}^n Y_i \qquad R_n = \frac{S_n}{n}.$$

de Finetti's Representation

Theorem: 0-1 representation theorem

We define

$$\theta_0 = \lim_{n \rightarrow \infty} R_n$$

that is, $R_n \xrightarrow{\text{a.s.}} \theta_0$.

The quantity θ_0 is the limiting relative frequency of 1s in the infinitely exchangeable binary sequence.

Proof: See Handout 01.

de Finetti's Representation

Note

- (i) The converse of the theorem is also true: it is straightforward to see that the distributions formed by computing the integral for a given $\pi_0(\cdot)$ are finite dimensional distributions derived for an infinitely exchangeable sequence.

Note

- (ii) The quantity θ parameterizes the conditional distribution of the Y_i ; we can interpret

$$\prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} = \prod_{i=1}^n p_{Y_i}(y_i; \theta)$$

and deduce that for each n , Y_1, \dots, Y_n are *conditionally independent* given θ .

Note

(iii) θ_0 is the true *data-generating* value, and under this model

$$Y_1, \dots, Y_n \sim \text{Bernoulli}(\theta_0)$$

are *independent*.

θ denotes a *generic* possible value of the parameter.

de Finetti's Representation

Note

- (iv) $\pi_0(\cdot)$ is a probability distribution for θ , but the generality of the construction *does not specify* what form $\pi_0(d\theta)$ should take; different choices for $\pi_0(\cdot)$ will lead to different exchangeable forms.
- (v) We often relax notation and allow $\pi_0(\cdot)$ to denote either the cdf or the pdf whenever convenient to do so.

The theorem extends to *arbitrary* infinitely exchangeable sequences.

Theorem: General representation theorem

Suppose that

- Y_1, Y_2, \dots is an infinitely exchangeable sequence of variables taking values on \mathbb{R} ;
- P_Y is a probability measure on \mathbb{R}^∞ that defines all finite dimensional distributions for $\{Y_n\}_{n=1}^\infty$;
- \mathcal{F} denotes the set of all distribution functions on \mathbb{R} .

de Finetti's Representation

Theorem: General representation theorem

Then there exists a distribution function $\pi_0(\cdot)$ on \mathcal{F} , such that the joint distribution of (Y_1, Y_2, \dots, Y_n) has the form

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \int_{\mathcal{F}} \left\{ \prod_{i=1}^n F(y_i) \right\} \pi_0(dF)$$

where F parameterizes the model: F is an unobservable distribution function.

de Finetti's Representation

Theorem: General representation theorem

We interpret F via its limiting form; let F_0 be a distribution function defined for $y \in \mathbb{R}$ by

$$F_0(y) = \lim_{n \rightarrow \infty} F_n(y) = \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(Y_i) \right\}$$

is a distribution on the space of functions \mathcal{F} , defined as a limit as $n \rightarrow \infty$ of the *empirical distribution function*, F_n , defined for Y_1, \dots, Y_n .

Note

- (i) The unknown distribution F parameterizes the conditional distribution of the Y_i ;

$$\prod_{i=1}^n F(y_i)$$

indicates that for each n , Y_1, \dots, Y_n are *conditionally independent* given F .

Note

(ii) F_0 is the *limiting empirical distribution function*:

- classical results tell us that this limiting distribution can be interpreted as the *true* marginal distribution function for the Y_i ;
- the limiting form does not tell us about the joint structure of the Y_i .

Note

(iii) $\pi_0(\cdot)$ is a probability distribution for F ; therefore it is a probability distribution on the *space \mathcal{F} of distribution functions*.

\mathcal{F} is a more complicated space than \mathbb{R}^d , say, but the procedure for specifying a distribution on \mathcal{F} remains essentially the same:

- we consider subsets of B of \mathcal{F} , and then assign $\Pr(F \in B)$;
- calculations require (Lebesgue) integrals over \mathcal{F} taken with respect to π_0

Note

(iv) F_n is the *empirical distribution function*:

- this is the classical estimator of the distribution function based on Y_1, \dots, Y_n ;
- pointwise behaviour (at each individual y) easy to study;
- function-wise behaviour (at all y simultaneously) requires *empirical process theory*.

de Finetti's Representation

The general representation theorem can be made specific by

- imposing *symmetry* or *invariance constraints* on the observables;
- requiring the existence of *sufficient statistics*;
 - the *exponential family*.
- allowing for *partial exchangeability* to construct conditional forms of exchangeable sequences
 - *regression*, *hierarchical models* etc.

These considerations lead to the use of specific *parametric* models.

Example: Normal model

- $\mathcal{Y} \equiv \mathbb{R}$;
- $\theta = (\mu, \sigma)$;
- $f_Y(y; \theta) \equiv \text{Normal}(\mu, \sigma^2)$
- $\pi_0(\mu, \sigma)$ a prior density on $\mathbb{R} \times \mathbb{R}^+$.

de Finetti's Representation

Example: Normal model

This model can be justified by a requirement for *centered spherical symmetry* amongst the Y variables, that is, for all $n \geq 1$ the quantities

$$Y_1 - \bar{Y}_n, \dots, Y_n - \bar{Y}_n$$

exhibit *spherical symmetry*, that is, if $Z_i = Y_i - \bar{Y}_n$ for $i = 1, \dots, n$, then the vectors

$$\mathbf{Z}_n = (Z_1, \dots, Z_n)^\top \quad \text{and} \quad \mathbf{A}\mathbf{Z}_n$$

have the same distribution, for any $n \times n$ matrix \mathbf{A} such that $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_n$.

de Finetti's Representation

Example: Normal model

The de Finetti representation then gives the joint density of the observables as

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \int_{-\infty}^{\infty} \int_0^{\infty} \left\{ \prod_{i=1}^n f_Y(y_i; \mu, \sigma) \right\} \pi_0(d\mu, d\sigma)$$

for $n \geq 1$.

de Finetti's Representation

A typical form of the de Finetti representation is in terms of parametric densities: for $n \geq 1$

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \int \prod_{i=1}^n f(y_i; \theta) \pi_0(d\theta)$$

where

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$$

is the *joint pdf* for Y_1, \dots, Y_n .

de Finetti's Representation

Note

The de Finetti calculation is a type of '*marginalization*' calculation; for example, for two continuous random variables

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

We can think of $f_X(x)$ as a '*mixing*' distribution.

In the de Finetti representation, the two random variables are

- the observables Y_1, \dots, Y_n ;
- the 'parameter' θ or F .

de Finetti's Representation

Note

The terms

$$\prod_{i=1}^n p_{Y_i}(y_i; \theta) \quad \text{or} \quad \prod_{i=1}^n f_{Y_i}(y_i; \theta)$$

in parametric case, or in the non-parametric case, say

$$\prod_{i=1}^n dF(y_i)$$

are equivalent to the familiar *likelihood function* that forms the basis of much statistical theory.

Prediction

The assumption of infinite exchangeability and the de Finetti representation give an automatic rule for constructing predictions. For $n, m \geq 1$ consider the prediction of

$$Y_{n+1}, \dots, Y_{n+m}$$

conditional on observed values of

$$Y_1, \dots, Y_n.$$

We focus first on the binary case for simplicity.

Prediction

By de Finetti, recall that for each $n \geq 1$

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \int_0^1 \left\{ \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \pi_0(d\theta) \quad (\S)$$

Prediction

Similarly

$$p_{Y_1, \dots, Y_{n+m}}(y_1, \dots, y_{n+m}) = \int_0^1 \left\{ \prod_{i=1}^{n+m} \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \pi_0(d\theta)$$

Prediction

Then for the *predictive* distribution, by the conditional probability definition, we have

$$\begin{aligned} p_{Y_{n+1}, \dots, Y_{n+m} | Y_1, \dots, Y_n} (y_{n+1}, \dots, y_{n+m} | y_1, \dots, y_n) \\ &= \frac{p_{Y_1, \dots, Y_{n+m}} (y_1, \dots, y_{n+m})}{p_{Y_1, \dots, Y_n} (y_1, \dots, y_n)} \\ &= \frac{\int_0^1 \left\{ \prod_{i=1}^{n+m} \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \pi_0 (d\theta)}{\int_0^1 \left\{ \prod_{i=1}^n t^{y_i} (1 - t)^{1-y_i} \right\} \pi_0 (dt)} \end{aligned}$$

where t is a dummy integrating variable.

Prediction

We may rewrite this expression by noting that the denominator can be treated as a constant in the integral in the numerator, and that the product in the numerator can be split

$$\prod_{i=1}^{n+m} \theta^{y_i} (1 - \theta)^{1-y_i} \\ = \left\{ \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \times \left\{ \prod_{i=n+1}^{n+m} \theta^{y_i} (1 - \theta)^{1-y_i} \right\}.$$

Prediction

That is

$$\begin{aligned} p_{Y_{n+1}, \dots, Y_{n+m} | Y_1, \dots, Y_n} (y_{n+1}, \dots, y_{n+m} | y_1, \dots, y_n) \\ = \int_0^1 \left\{ \prod_{i=n+1}^{n+m} \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \pi_n (d\theta) \end{aligned} \quad (\dagger)$$

where

$$\pi_n (d\theta) = \frac{\prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \pi_0 (d\theta)}{\int_0^1 \left\{ \prod_{i=1}^n t^{y_i} (1 - t)^{1-y_i} \right\} \pi_0 (dt)} \quad (\ddagger)$$

Prediction

Comparing (§) to (†), we see that the forms of the two representations for

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$$

and

$$p_{Y_{n+1}, \dots, Y_{n+m} | Y_1, \dots, Y_n}(y_{n+1}, \dots, y_{n+m} | y_1, \dots, y_n)$$

are *identical* with $\pi_0(d\theta)$ in the former replaced by $\pi_n(d\theta)$ in the latter.

Prediction

We can therefore think of

$$\pi_n(d\theta)$$

as being an *updated* version of

$$\pi_0(d\theta)$$

in light of observing y_1, \dots, y_n . Note that

$$\int \pi_n(d\theta) = 1$$

from (‡), so $\pi_n(d\theta)$ does define a *valid probability distribution*.

Terminology

- $\pi_0(d\theta)$ is the *prior distribution* for θ ;
- $\pi_n(d\theta)$ is the *posterior distribution* for θ ;
- $p_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$ is the *prior predictive distribution*
 - also termed the *marginal likelihood*;
- $p_{Y_{n+1}, \dots, Y_{n+m} | Y_1, \dots, Y_n}(y_{n+1}, \dots, y_{n+m} | y_1, \dots, y_n)$ is the *posterior predictive distribution*.

Limiting predictions

Let

$$S_{1,n} = \sum_{i=1}^n Y_i \qquad S_{n+1,n+m} = \sum_{i=n+1}^{n+m} Y_i$$

By direct calculation from (§), by the theorem of total probability, we have

$$\Pr[S_{1,n} = s_{1,n}] = \binom{n}{s_{1,n}} \int_0^1 t^{s_{1,n}} (1-t)^{n-s_{1,n}} \pi_0(dt)$$

using t as the integrating variable; this holds for

$$s_{1,n} \in \{0, 1, \dots, n\}.$$

Limiting predictions

This follows as, by exchangeability, all sequences containing $s_{1,n}$ ones have the same probability, namely,

$$\int_0^1 t^{s_{1,n}} (1-t)^{n-s_{1,n}} \pi_0(dt),$$

and there are

$$\binom{n}{s_{1,n}}$$

such sequences.

Limiting predictions

Note from (‡) that $\pi_n(d\theta)$ depends on the data y_1, \dots, y_n only via

$$s_{1,n} = \sum_{i=1}^n y_i$$

as

$$\pi_n(d\theta) = \frac{\theta^{s_{1,n}} (1 - \theta)^{n-s_{1,n}} \pi_0(d\theta)}{\int_0^1 t^{s_{1,n}} (1 - t)^{n-s_{1,n}} \pi_0(dt)}$$

Thus we can interpret $S_{1,n}$ as a (Bayesian) *sufficient statistic*.

Limiting predictions

Therefore for $s \in \{0, 1, \dots, m\}$,

$$\begin{aligned}\Pr[S_{n+1,n+m} = s | S_{1,n} = s_{1,n}] \\ = \binom{m}{s} \int_0^1 t^s (1-t)^{m-s} \pi_n(dt).\end{aligned}$$

Limiting predictions

Now let

$$R_{n+1,n+m} = \frac{S_{n+1,n+m}}{m}.$$

Then, by the form of (\dagger), and the result from the theorem (\blacklozenge), we may conclude directly that

$$\lim_{m \rightarrow \infty} \Pr [R_{n+1,n+m} \leq \theta | S_{1,n} = s_{1,n}] = \int_0^\theta \pi_n(dt)$$

that is, the *posterior distribution* is a *limiting form of the predictive distribution* for a particular summary statistic.

Limiting predictions

Note

In the above formulation, if we consider $n \rightarrow \infty$, we observe that for $\theta \in [0, 1]$,

$$\lim_{n \rightarrow \infty} \pi_n(\theta) = \delta_{\{\theta_0\}}(\theta) = \begin{cases} 1 & \theta = \theta_0 \\ 0 & \theta \neq \theta_0 \end{cases}$$

that is, the posterior distribution is *degenerate* at θ_0 .

This happens provided that θ_0 lies within the support of the prior.

True values

In our specifications, we have defined

- in the binary case

$$\theta_0 = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i$$

- in the general case

$$F_0(y) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(Y_i)$$

We can regard θ_0 and $F_0(\cdot)$ as the ‘*true*’ parameters that render the observables independent if conditioned upon.

Correct specification

In any real example, we observe y_1, \dots, y_n , but whereas we can *propose* a model for the joint distribution of the variables under exchangeability, we do not know that our selected model is the correct (data generating) one.

The limit results hold under the assumption that the model is *correctly specified*, but in reality our model may be *mis-specified*.

Correct specification

Example: Binary case

In the exchangeable binary case, we must have

$$p_{Y_i}(y_i; \theta) = \theta^{y_i} (1 - \theta)^{1-y_i}$$

for each $0 \leq \theta \leq 1$; however different choices of $\pi_0(d\theta)$ lead to different models for the joint distribution of Y_1, \dots, Y_n .

θ_0 is the hypothetical value of θ that renders the Y_i s independent if conditioned upon: however,

- this *cannot be assessed in the data*, as we do not observe data from a conditional-on- θ model.

Correct specification

Example: Binary case

Suppose that

$$\pi_0(\theta) \equiv \text{Beta}(\alpha_0, \beta_0).$$

for $\alpha_0, \beta_0 > 0$. Then from (‡), we have for the posterior density

$$\begin{aligned}\pi_n(\theta) &\propto \left\{ \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} \right\} \theta^{\alpha_0-1} (1 - \theta)^{\beta_0-1} \\ &= \theta^{s_n + \alpha_0 - 1} (1 - \theta)^{n - s_n + \beta_0 - 1}\end{aligned}$$

where

$$s_n = \sum_{i=1}^n y_i.$$

Correct specification

Example: Binary case

That is,

$$\pi_n(\theta) \equiv \text{Beta}(s_n + \alpha_0, n - s_n + \beta_0) \equiv \text{Beta}(\alpha_n, \beta_n)$$

say, so that

$$\pi_n(\theta) = \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \theta^{\alpha_n-1} (1 - \theta)^{\beta_n-1}.$$

Correct specification

Example: Binary case

From inspection of posterior $\pi_n(\theta)$, we may deduce that

$$\lim_{n \rightarrow \infty} \pi_n(\theta) \longrightarrow \delta_{\theta_0}(\theta)$$

that is, the posterior is *degenerate* at θ_0 *irrespective* of the choice of prior parameters.

Correct specification

Example: Binary case

However, other choices of prior *may* lead to different behaviour: for example,

- if $\pi_0(\theta)$ is itself *degenerate* at a given value c say, then the posterior is also degenerate at c ;
- if $\pi_0(\theta)$ is *uniform on a sub-interval* $(c_1, c_2) \subset [0, 1]$, then the posterior is also restricted to this interval.

Correct specification

Example: Binary case

From (§), we may conclude that the prior predictive

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$$

takes the form

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_n)\Gamma(\beta_n)}{\Gamma(\alpha_n + \beta_n)}$$

Correct specification

Example: Binary case

For $n = 1$:

$$p_{Y_1}(y_1) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0 + y_1)\Gamma(\beta_0 + 1 - y_1)}{\Gamma(\alpha_0 + \beta_0 + 1)}$$

so that

$$\begin{aligned}\mathbb{E}_{Y_1}[Y_1] &= \Pr[Y_1 = 1] = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0 + 1)\Gamma(\beta_0)}{\Gamma(\alpha_0 + \beta_0 + 1)} \\ &= \frac{\alpha_0}{\alpha_0 + \beta_0}\end{aligned}$$

Correct specification

Example: Binary case

For $n = 2$:

$$p_{Y_1, Y_2}(y_1, y_2) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0 + y_1 + y_2)\Gamma(\beta_0 + 2 - y_1 - y_2)}{\Gamma(\alpha_0 + \beta_0 + 2)}$$

$$\begin{aligned}\therefore \mathbb{E}_{Y_1, Y_2}[Y_1 Y_2] &= \Pr[Y_1 = 1, Y_2 = 1] \\ &= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0 + 2)\Gamma(\beta_0)}{\Gamma(\alpha_0 + \beta_0 + 2)} \\ &= \frac{\alpha_0(\alpha_0 + 1)}{(\alpha_0 + \beta_0)(\alpha_0 + \beta_0 + 1)}\end{aligned}$$

Correct specification

Example: Binary case

$$\text{Cov}_{Y_1, Y_2}[Y_1, Y_2] = \frac{\alpha_0(\alpha_0 + 1)}{(\alpha_0 + \beta_0)(\alpha_0 + \beta_0 + 1)} - \left(\frac{\alpha_0}{\alpha_0 + \beta_0} \right)^2$$

Thus as the prior changes (that is, α_0 and β_0 change) the modelled covariance between pairs of Y s changes.

Correct specification

Example: Binary case

From (†), we may conclude that the posterior predictive

$$p_{Y_{n+1}, \dots, Y_{n+m} | Y_1, \dots, Y_n} (y_{n+1}, \dots, y_{n+m} | y_1, \dots, y_n)$$

takes the form

$$\frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \frac{\Gamma(\alpha_n + s_{n+1, n+m})\Gamma(\beta_n + m - s_{n+1, n+m})}{\Gamma(\alpha_n + \beta_n + m)}$$

where

$$s_{n+1, n+m} = \sum_{i=n+1}^{n+m} y_i$$

Correct specification

Example: Binary case

If $n = m = 1$

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{\Gamma(\alpha_0 + y_1 + y_2)\Gamma(\beta_0 + 2 - y_1 - y_2)}{\Gamma(\alpha_0 + y_1)\Gamma(\beta_0 + 1 - y_1)(\alpha_0 + \beta_0 + 1)}$$

so that for $y_2 \in \{0, 1\}$

$$p_{Y_2|Y_1}(y_2|0) = \frac{\Gamma(\alpha_0 + y_2)\Gamma(\beta_0 + 2 - y_2)}{\Gamma(\alpha_0)\Gamma(\beta_0 + 1)(\alpha_0 + \beta_0 + 1)}$$

$$p_{Y_2|Y_1}(y_2|1) = \frac{\Gamma(\alpha_0 + 1 + y_2)\Gamma(\beta_0 + 1 - y_2)}{\Gamma(\alpha_0 + 1)\Gamma(\beta_0)(\alpha_0 + \beta_0 + 1)}$$

Correct specification

In general, in the *parametric* case, we need to consider the possibility of mis-specification of a component of the model; the choice of

$$f_Y(y_i; \theta) \quad \text{or} \quad \pi_0(d\theta)$$

yields a prior predictive

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$$

that may not match the *true* (data generating) model.

In the *non-parametric* case, the same considerations apply.

Correct specification

De Finetti's theorem tells us that under exchangeability, there must exist a representation of the data generating model such that

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \int \prod_{i=1}^n F^*(y_i; \theta) \pi_0^*(d\theta)$$

for at least one combination of

$$F_Y^*(y_i; \theta) \quad \text{and} \quad \pi_0^*(d\theta)$$

Correct specification

It may be that $\pi_0^*(d\theta)$ is a degenerate distribution at $\theta = \theta_0^*$ say, so that

$$\Pr \left[\bigcap_{i=1}^n (Y_i \leq y_i) \right] = \prod_{i=1}^n F^*(y_i; \theta_0^*)$$

and the Y_i s are independent.

In most cases, we will assume correct specification.

Numerical Examples

Example: Normal model

See knitr handout 01.

Note the important identity: for scalar x , and constants A, a, B, b

$$A(x - a)^2 + B(x - b)^2 = (A + B) \left(x - \frac{Aa + Bb}{A + B} \right)^2 + \frac{AB}{A + B} (a - b)^2$$

Example: Bernoulli model

See knitr handout 02.

Extensions

- multivariate Y s: extension straightforward;
- regression problems: *partial exchangeability*;
- hierarchical models: *partial exchangeability*.

Bayesian Calculations & Specifications

Bayesian calculations

- Bayesian updating
- Sufficiency concepts
- Prior specification

Bayesian updating

The Bayesian calculation acts *sequentially*; that is, for data \mathbf{y}_1

$$\pi_{n_1}(\theta) = \frac{f_{\mathbf{Y}_1}(\mathbf{y}_1; \theta)\pi_0(\theta)}{f_{\mathbf{Y}_1}(\mathbf{y}_1)} = \frac{f_{\mathbf{Y}_1}(\mathbf{y}_1; \theta)\pi_0(\theta)}{\int f_{\mathbf{Y}_1}(\mathbf{y}_1; t)\pi_0(t) dt}$$

contains the information about θ in light of the data \mathbf{y}_1 and prior assumptions.

Bayesian updating

If additional (exchangeable) data \mathbf{y}_2 become available, then the posterior for θ in light of the combined data $(\mathbf{y}_1, \mathbf{y}_2)$ is

$$\pi_n(\theta) = \frac{f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2; \theta) \pi_0(\theta)}{f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)} = \frac{f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2; \theta) \pi_0(\theta)}{\int f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2; t) \pi_0(t) dt}$$

where $n = n_1 + n_2$ is the total sample size.

Bayesian updating

But note also that

$$\pi_n(\theta) = \frac{f_{\mathbf{Y}_2}(\mathbf{y}_2; \theta) \pi_{n_1}(\theta)}{f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2|\mathbf{y}_1)}$$

where $\pi_{n_1}(\theta)$ is the posterior for θ based on \mathbf{y}_1 , and

$$f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2|\mathbf{y}_1) = \frac{f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)}{f_{\mathbf{Y}_1}(\mathbf{y}_1)} = \frac{\int f_{\mathbf{Y}_1, \mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2; t) \pi_0(t) dt}{\int f_{\mathbf{Y}_1}(\mathbf{y}_1; s) \pi_0(s) ds}$$

Sufficiency

Let

$$\mathbf{Y} = (Y_1, \dots, Y_n) \quad \mathbf{y} = (y_1, \dots, y_n)$$

and consider parametric model $f_Y(y; \theta)$.

$\mathbf{T}(\mathbf{Y})$ is a *sufficient statistic* for θ in the classical sense if we can write

$$f_Y(\mathbf{y}; \theta) = g(\mathbf{T}(\mathbf{y}), \theta)h(\mathbf{y}).$$

for some functions g and h .

Sufficiency

That is, the dependence of the joint pdf on θ is encapsulated in a function that depends on the value of the *summary statistic*

$$\mathbf{T}(\mathbf{y}) = \mathbf{T}(y_1, \dots, y_n)$$

and *not* on the individual y values.

Sufficiency

It follows for the posterior distribution that

$$\begin{aligned}\pi_n(\theta) &= \frac{f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi_0(\theta)}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{g(\mathbf{T}(\mathbf{y}), \theta)h(\mathbf{y})\pi_0(\theta)}{f_{\mathbf{Y}}(\mathbf{y})} \\ &= \left[\frac{h(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} \right] g(\mathbf{T}(\mathbf{y}), \theta)\pi_0(\theta)\end{aligned}$$

Thus the posterior distribution of θ only depends on the data through $\mathbf{T}(\mathbf{y})$.

Sufficiency

Lemma

If $\mathbf{T}(\mathbf{Y})$ is a sufficient statistic for θ (in the classical sense) then $\pi_n(\theta)$ depends on \mathbf{y} only through the value of

$$\mathbf{T}(\mathbf{y})$$

for all prior specifications $\pi_0(\theta)$.

Sufficiency

Proof.

By definition

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = f_{\mathbf{Y}, \mathbf{T}}(\mathbf{y}, \mathbf{t}; \theta)$$

if $\mathbf{t} = \mathbf{T}(\mathbf{y})$, and zero otherwise. Thus, by sufficiency,

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = f_{\mathbf{Y}|\mathbf{T}}(\mathbf{y}|\mathbf{t})f_{\mathbf{T}}(\mathbf{t}; \theta)$$

and hence

$$\pi_n(\theta) \propto f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi(\theta) \propto f_{\mathbf{T}}(\mathbf{t}; \theta)\pi_0(\theta)$$



Sufficiency

Lemma

$\mathbf{T}(\mathbf{Y})$ is sufficient in the Bayesian sense – the Bayesian posterior depends on \mathbf{y} only through $\mathbf{T}(\mathbf{y})$ – if and only if it is sufficient in the classical sense.

We need to establish the converse of the previous result.

Sufficiency

Proof.

For the posterior based on \mathbf{t} ,

$$\pi_n(\theta) = \frac{f_{\mathbf{T}}(\mathbf{t}; \theta) \pi_0(\theta)}{f_{\mathbf{T}}(\mathbf{t})}.$$

This must be equal to the posterior based on \mathbf{Y} , that is,

$$\frac{f_{\mathbf{Y}}(\mathbf{y}; \theta)}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{\pi_n(\theta)}{\pi_0(\theta)} = \frac{f_{\mathbf{T}}(\mathbf{t}; \theta)}{f_{\mathbf{T}}(\mathbf{t})}.$$

Hence we must have

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = f_{\mathbf{T}}(\mathbf{t}; \theta) \frac{f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{T}}(\mathbf{t})} = g(\mathbf{t}, \theta) h(\mathbf{y})$$

say. Thus $\mathbf{T}(\mathbf{Y})$ is sufficient in the classical sense. \square

Construction of Prior Distributions

In the Bayesian formulation, the prior density plays an important role. There are several methods via which the prior can be specified quantitatively;

- from *historical* or training data;
- by *subjective assessment* after quantitative *elicitation*;
- by matching to a *desired functional form*;
- or in a *non-informative* or *vague* specification, where the prior probability is supposedly spread ‘evenly’ across the parameter space.

Conjugate Priors

For some models, a *conjugate prior* can be chosen; this prior combines with the likelihood in such a way to give an analytically tractable posterior calculation.

Consider a class of distributions \mathcal{F} indexed by parameter θ

$$\mathcal{F} = \{f_Y(y; \theta) : \theta \in \Theta\}$$

A class \mathcal{P} of prior distributions for θ is a *conjugate family* for \mathcal{F} if the posterior distribution for θ resulting from data \mathbf{y} is an element of \mathcal{P} for all $f_Y \in \mathcal{F}$, $\pi_0 \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Y}$.

Conjugate Priors

Example: Exponential Family

Suppose that $f_Y(y; \theta)$ is an Exponential Family distribution

$$f_Y(y; \theta) = h(y)c(\theta) \exp \left\{ \sum_{j=1}^k t_j(y) w_j(\theta) \right\}$$

so that for a random sample of size n

$$\mathcal{L}_n(\theta) = h(\mathbf{y}) \{c(\theta)\}^n \exp \left\{ \sum_{j=1}^k T_j(\mathbf{y}) w_j(\theta) \right\} \quad (1)$$

for

$$T_j(\mathbf{y}) = \sum_{i=1}^n t_j(y_i).$$

Conjugate Priors

Example: Exponential Family

Suppose that

$$\pi_0(\theta) = d(\alpha, \beta) \{c(\theta)\}^\alpha \exp \left\{ \sum_{j=1}^k \beta_j w_j(\theta) \right\} \quad (2)$$

where α and $\beta = (\beta_1, \dots, \beta_k)^\top$ are *hyperparameters*.

Conjugate Priors

Example: Exponential Family

Combining prior and likelihood yields the posterior as

$$\begin{aligned}\pi_n(\theta) &\propto \{c(\theta)\}^{\alpha+n} \exp \left\{ \sum_{j=1}^k [\beta_j + T_j(\mathbf{y})] w_j(\theta) \right\} \\ &= \{c(\theta)\}^{\alpha^*} \exp \left\{ \sum_{j=1}^k \beta_j^* w_j(\theta) \right\}\end{aligned}$$

Conjugate Priors

Example: Exponential Family

The normalizing constant can be deduced to be

$$d(\alpha + n, \beta + \mathbf{T}(\mathbf{y})),$$

and hence the posterior distribution has the same functional form as the prior, but with parameters updated to

$$\alpha^* = \alpha + n \quad \beta^* = (\beta_1^*, \dots, \beta_k^*)^\top = (\beta_1 + T_1(\mathbf{y}), \dots, \beta_k + T_k(\mathbf{y}))^\top.$$

Ignorance Priors

A non-informative prior expresses *prior ignorance* about the parameter of interest.

- If $\Theta = \{\theta_1, \dots, \theta_k\}$ (that is, θ is known to take one of a finite number of possible values). Then a non-informative prior places equal probability on each value, that is,

$$\pi_0(\theta) = \frac{1}{k} \quad \theta \in \Theta.$$

- If Θ is a *bounded region*, then a natural non-informative prior is *constant* on Θ .

Ignorance Priors

- If the parameter space Θ is not finite, or is unbounded, however, a non-informative prior specification is more difficult to construct.
 - A naive prior specification would be to set $\pi_0(\theta)$ to be a *constant*; however this prior does not sum/integrate to 1 over Θ .
 - A prior distribution $\pi_0(\theta)$ for parameter θ is termed *improper* if it does not sum/integrate to 1.

Note that even for an improper prior, the posterior density may still be *proper* (integrate to 1).

Ignorance Priors

However, if $\phi = g(\theta)$ is a transformation of θ , then by elementary transformation results, including the Jacobian of the transform $J(\theta \rightarrow \phi)$, it follows that

$$\pi_{0,\theta}(\theta) = c \quad \implies \quad \pi_{0,\phi}(\phi) = c \times J(\theta \rightarrow \phi)$$

which may *not* be constant, and hence a *non-uniform* prior on ϕ results. This is perhaps unsatisfactory, and so the following procedure may be preferable.

The Jeffreys Prior

Consider the prior $\pi_0(\theta)$ for parameter θ in probability model $f_Y(y; \theta)$ determined by

$$\pi_0(\theta) \propto \{|\mathcal{I}_\theta(\theta)|\}^{1/2}$$

where $\mathcal{I}_\theta(\theta)$ is the *Fisher Information*,

$$\mathcal{I}_\theta(\theta) = \mathbb{E}_Y [S(Y; \theta) S(Y; \theta)^\top; \theta] = -\mathbb{E}_Y [\Psi(Y; \theta); \theta]$$

and $|\mathcal{I}_\theta(\theta)|$ indicates the absolute value of the determinant of $\mathcal{I}_\theta(\theta)$. The prior $\pi_0(\theta)$ defined in this way is termed the *Jeffreys* prior.

The Jeffreys Prior

$S(y; \theta)$ is the $k \times 1$ vector *score function* with j th element

$$S_j(y; \theta) = \frac{\partial}{\partial \theta_j} \log f_Y(y; \theta) \quad j = 1, \dots, k$$

and $\Psi(Y; \theta)$ is the $k \times k$ matrix of second partial derivatives with (j, l) th element

$$\frac{\partial^2}{\partial \theta_j \partial \theta_l} \log f_Y(y; \theta)$$

Note that $S(Y; \theta)$ and $\Psi(Y; \theta)$ are *random variables*.

The Jeffreys Prior

Example: *Binomial*(m, θ)

We have

$$\log f_Y(y; \theta) = \log \binom{m}{y} + y \log \theta + (m - y) \log(1 - \theta)$$

$$S(y; \theta) = \frac{y}{\theta} - \frac{(m - y)}{(1 - \theta)}$$

$$\Psi(y; \theta) = -\frac{y}{\theta^2} - \frac{(m - y)}{(1 - \theta)^2}$$

The Jeffreys Prior

Example: *Binomial*(m, θ)

Therefore

$$\mathcal{I}_{\theta}(\theta) = -\mathbb{E}_Y \left[-\frac{Y}{\theta^2} - \frac{(m - Y)}{(1 - \theta)^2}; \theta \right] = \frac{m\theta}{\theta^2} + \frac{m(1 - \theta)}{(1 - \theta)^2} = \frac{m}{\theta(1 - \theta)}$$

and hence

$$\pi_0(\theta) \propto |\mathcal{I}_{\theta}(\theta)|^{1/2} = \{\theta(1 - \theta)\}^{-1/2}$$

The Jeffreys Prior

Lemma

Jeffreys's prior is invariant under 1-1 transformations, that is, if $\phi = \phi(\theta)$, then the prior for ϕ obtained by reparameterization from θ to ϕ in the prior for θ , is precisely Jeffreys's prior for ϕ .

The Jeffreys Prior

Proof.

Let $\phi = \phi(\theta)$ be a 1-1 transformation. Denote by $\ell_\theta(y; \theta)$ and $\ell_\phi(y; \phi)$ the log pdfs in the two parameterizations. Then by the rules of partial differentiation

$$\frac{\partial \ell_\phi}{\partial \phi_j} = \sum_{l=1}^k \frac{\partial \ell_\theta}{\partial \theta_l} \frac{\partial \theta_l}{\partial \phi_j} \quad j = 1, \dots, k$$

so that

$$S(y; \phi) = \Lambda(\theta, \phi) S(y; \theta)$$

where $\Lambda(\theta, \phi)$ is the $k \times k$ matrix with (j, l) th element

$$\frac{\partial \theta_l}{\partial \phi_j}$$

□

The Jeffreys Prior

Proof.

In fact, $\Lambda(\theta, \phi)$ is just the Jacobian of the transformation from θ to ϕ , $J(\theta \rightarrow \phi)$. Hence

$$\mathcal{I}_\phi(\phi) = \Lambda(\theta, \phi) \mathcal{I}_\theta(\theta) \Lambda(\theta, \phi)^\top$$

and so

$$|\mathcal{I}_\phi(\phi)| = |\Lambda(\theta, \phi) \mathcal{I}_\theta(\theta) \Lambda(\theta, \phi)^\top| = |\Lambda(\theta, \phi)|^2 |\mathcal{I}_\theta(\theta)|$$

and

$$|\mathcal{I}_\phi(\phi)|^{1/2} = |\Lambda(\theta, \phi)| |\mathcal{I}_\theta(\theta)|^{1/2}.$$



The Jeffreys Prior

Proof.

Thus

$$\pi_0(\phi) \propto |\mathcal{I}_\phi(\phi)|^{1/2} = |\Lambda(\theta, \phi)| |\mathcal{I}_\theta(\theta)|^{1/2} = |\Lambda(\theta, \phi)| \pi_0(\theta)$$

and Jeffreys's prior for ϕ is identical to the one that would be obtained by constructing Jeffreys's prior for θ and reparameterizing to ϕ . □

The Jeffreys Prior

Example: *Binomial*(m, θ)

Suppose that $\phi = \theta/(1 - \theta)$ (so that $\theta = \phi/(1 + \phi)$). Then

$$\log f_Y(y; \phi) = \log \binom{m}{y} + y \log \phi - m \log(1 + \phi)$$

$$S(y; \phi) = \frac{y}{\phi} - \frac{m}{(1 + \phi)}$$

$$\Psi(y; \phi) = -\frac{y}{\phi^2} + \frac{m}{(1 + \phi)^2}$$

The Jeffreys Prior

Example: *Binomial*(m, θ)

Therefore

$$\begin{aligned}\mathcal{I}_\phi(\phi) &= -\mathbb{E}_Y \left[-\frac{Y}{\phi^2} + \frac{m}{(1+\phi)^2}; \phi \right] \\ &= \frac{m\phi}{(1+\phi)\phi^2} - \frac{m}{(1+\phi)^2} \\ &= \frac{m}{\phi(1+\phi)^2}\end{aligned}$$

and hence

$$\pi_0(\phi) \propto |\mathcal{I}_\phi(\phi)|^{1/2} = \{\phi(1+\phi)^2\}^{-1/2}.$$

The Jeffreys Prior

Example: *Binomial*(m, θ)

Now, recall that Jeffreys's prior for θ takes the form

$$\pi_0(\theta) \propto \{\theta(1 - \theta)\}^{-1/2}$$

The Jacobian of the transformation from θ to ϕ is $(1 + \phi)^{-2}$, and thus using the univariate transformation theorem

$$\pi_0(\phi) \propto \{\phi/(1 + \phi)^2\}^{-1/2} (1 + \phi)^{-2} = \{\phi(1 + \phi)^2\}^{-1/2}$$

matching the result found above.

Location and Scale Parameters

Parameter θ is a *location parameter* if

$$f_Y(y; \theta) = f(y - \theta)$$

and is a *scale parameter* if

$$f_Y(y; \theta) = \frac{1}{\theta} f\left(\frac{y}{\theta}\right)$$

for some pdf f .

A ‘non-informative’ prior can be constructed using invariance principles in the location and scale cases.

Location and Scale Parameters

For a *location* parameter, for a non-informative prior, it is required to have, for set $A \subset \Theta$

$$\int_A \pi_0(\theta) d\theta = \int_{A_c} \pi_0(\theta) d\theta$$

where $A_c = \{\theta : \theta - c \in A\}$ for scalar c . Therefore, for all c , we must have

$$\int_{A_c} \pi_0(\theta) d\theta = \int_A \pi_0(\theta - c) d\theta$$

$$\therefore \pi_0(\theta) = \pi_0(\theta - c) \implies \pi_0(\theta) = \text{constant}.$$

Location and Scale Parameters

For a *scale* parameter, it is required to have, for arbitrary set $A \subset \Theta$

$$\int_A \pi_0(\theta) d\theta = \int_{A_c} \pi_0(\theta) d\theta$$

where now $A_c = \{\theta : c\theta \in A\}$ for scalar c . Therefore, for all c , we must have

$$\int_{A_c} \pi_0(\theta) d\theta = \int_A c\pi_0(c\theta) d\theta$$

$$\therefore \pi_0(\theta) = c\pi_0(c\theta) \implies \pi_0(\theta) \propto \frac{1}{\theta}$$

Location and Scale Parameters

This follows by the usual ‘scale invariance’ definition: a function $g(y)$ is *scale invariant* if

$$g(cy) \propto g(y)$$

and all scale invariant functions are power laws; for some $\alpha > 0$,

$$g(y) \propto y^{-\alpha}.$$

Here, the condition $\pi_0(\theta) = c\pi_0(c\theta)$ means that we must have $\alpha = 1$.

Optimal Decision Making

Bayesian Optimal Decisions

Many statistical procedures involve *decision-making*, that is, taking actions in light of observed data.

- parameter estimation;
- hypothesis testing;
- prediction/classification;
- model selection.

Bayesian Optimal Decisions

Define

- $T(\cdot)$ as a function of data $\mathbf{Y} = (Y_1, \dots, Y_n)$;

$$T : \mathbb{R}^n \longrightarrow \mathcal{T}$$

For example

$$T(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{sample mean}$$

$$T(\mathbf{Y}) = (Y_{(1)}, \dots, Y_{(n)})^\top \quad \text{order statistics}$$

$$T_Y(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(Y_i) \quad \text{empirical cdf}$$

- Model space \mathcal{F} ;

Bayesian Optimal Decisions

- Loss function, $L(.,.)$,

$$L : \mathcal{T} \times \mathcal{F} \longrightarrow \mathbb{R}^+ \bigcup \{0\}.$$

Defines the loss in reporting T when the truth is defined by $F \in \mathcal{F}$.

Example:

For cdf F_Y , let

$$\mu = \int y F_Y(dy).$$

Then could define

$$L(T, F_Y) = (T - \mu)^2$$

as the loss in reporting 'estimator' T when the true functional of interest is μ .

Bayesian Optimal Decisions

The *optimal decision* is one that minimizes the expected loss, where the expectation is taken with respect to the distribution of random quantities in the calculation.

For a parametric analysis parameterized by θ

- in a *frequentist* analysis, θ is a fixed constant and the data are treated as random;
- in a *Bayesian* analysis, the data y_1, \dots, y_n are fixed, and θ is a random variable.

Bayesian Optimal Decisions

Example: Frequentist calculation

For cdf F_Y , let

$$\mu = \int y F_Y(dy).$$

with

$$L(T, F_Y) = (T - \mu)^2$$

we have that

$$\begin{aligned} \arg \min_T \mathbb{E}_{F_Y}[(T - \mu)^2] \\ &= \arg \min_T \{ \mathbb{E}_{F_Y}[(T - \mathbb{E}_{F_Y}[T])^2] + (\mathbb{E}_{F_Y}[T] - \mu)^2 \} \\ &= \arg \min_T \{ \text{Var}_{F_Y}[T] + (\mathbb{E}_{F_Y}[T] - \mu)^2 \} \end{aligned}$$

Example: Frequentist calculation

This does not define the optimal T , but it does tell us that we need to take into account

- the **variance** of T , $\text{Var}_{F_Y}[T]$
- the squared **bias**, $b_{F_Y}(T)$

$$b_{F_Y}(T) = \mathbb{E}_{F_Y}[T] - \mu$$

Kullback-Leibler loss

The *Kullback-Leibler* (KL) loss is used when measuring the discrepancy between distributions. For two distributions with cdfs F_0, F_1

$$KL(F_0, F_1) = \int \log \left\{ \frac{dF_0(y)}{dF_1(y)} \right\} dF_0(y)$$

which is defined when F_1 is absolutely continuous with respect to F_0 , that is for the corresponding probability measures

$$P_0(B) = 0 \implies P_1(B) = 0$$

for any set B .

Kullback-Leibler loss

- Discrete case:

$$KL(p_0, p_1) = \sum_y \log \left\{ \frac{p_0(y)}{p_1(y)} \right\} p_0(y) = \mathbb{E}_{p_0} \left[\log \left\{ \frac{p_0(Y)}{p_1(Y)} \right\} \right].$$

- Continuous case:

$$KL(f_0, f_1) = \int \log \left\{ \frac{f_0(y)}{f_1(y)} \right\} f_0(y) \, dy = \mathbb{E}_{f_0} \left[\log \left\{ \frac{f_0(Y)}{f_1(Y)} \right\} \right].$$

Kullback-Leibler loss

Note

1. $KL(F_0, F_1) \geq 0$;
2. $KL(F_0, F_1) \neq KL(F_1, F_0)$ in general;
3. $KL(F_0, F_1) = 0$ if and only if the two distributions are identical.

Kullback-Leibler loss

Example:

In a parametric problem, we might have pdf

$$f(y; \theta)$$

with $\theta = \theta_0$ presumed to be the data generating model. Then we may write

$$KL(\theta_0, \theta) = \int \log \left\{ \frac{f(y; \theta_0)}{f(y; \theta)} \right\} f(y; \theta_0) dy$$

and we seek to use data to report an estimator $\hat{\theta} = T(Y_{1:n})$ of the true value θ_0 .

Decision theory concepts

The key components of a *decision problem* are as follows;

- a *decision* d is to be made, and the decision is selected from some set \mathcal{D} of alternatives.
- a true *state of nature*, $v(\theta)$, lying in set Υ , defined by the data generating model, $F_Y(y; \theta)$.
- a *loss function*, $L(d, v)$, for decision d and state v , which records the loss (or penalty) incurred when the true state of nature is v and the decision made is d .

We aim to select the decision to *minimize* the *expected loss*.

Decision theory concepts

In an *estimation* context, the decision required is to *estimate* the parameter, and the true state of nature is the value of the parameter, $v(\theta) \equiv \theta$.

If data $\mathbf{y} = (y_1, \dots, y_n)$ are available, the optimal decision will in general be a function of the data.

Decision theory concepts

Suppose now that the ‘decision’ is in the form of a statistic, that summarizes the observed data, say

$$\mathbf{d}(\mathbf{y}) \equiv T(\mathbf{y}) = t_n.$$

with associated loss $L(t_n, \theta)$. The corresponding random variable version is

$$T_n \equiv T(\mathbf{Y}).$$

Decision theory concepts

The *frequentist risk* or *loss* associated with decision denoted $d(\mathbf{Y})$ is the expected loss associated with $d(\mathbf{Y})$, with the expectation taken over the distribution of \mathbf{Y} given θ

$$R_n(d, \theta) = \mathbb{E}_{F_{\mathbf{Y}}}[L(T_n, \theta)] = \int_{\mathcal{Y}} L(T(\mathbf{y}), \theta) f_{\mathbf{Y}}(\mathbf{y}; \theta) d\mathbf{y}$$

Decision theory concepts

The (*average*) *Bayes risk* for $d(\mathbf{Y})$ is the expected risk over the *prior* distribution of θ :

$$\begin{aligned} R_n(d) &= \mathbb{E}_{\pi_0}[R_n(d, \theta)] \\ &= \mathbb{E}_{\pi_0} [\mathbb{E}_{F_Y} [L(T_n, \theta)]] \\ &= \int_{\Theta} \left\{ \int_{\mathcal{Y}} L(t_n, \theta) f_Y(\mathbf{y}; \theta) d\mathbf{y} \right\} \pi_0(\theta) d\theta \quad t_n \equiv t_n(\mathbf{y}) \\ &= \int_{\Theta} \int_{\mathcal{Y}} L(t_n, \theta) f_Y(\mathbf{y}) \pi_n(\theta) d\mathbf{y} d\theta \\ &= \int_{\mathcal{Y}} \left\{ \int_{\Theta} L(t_n, \theta) \pi_n(\theta) d\theta \right\} f_Y(\mathbf{y}) d\mathbf{y} \end{aligned}$$

where by Bayes theorem $f_Y(\mathbf{y}; \theta) \pi_0(\theta) = f_Y(\mathbf{y}) \pi_n(\theta)$.

Decision theory concepts

With prior $\pi_0(\theta)$ and fixed data \mathbf{y} the optimal Bayesian decision, termed the *Bayes rule* is

$$\hat{\mathbf{d}}_B = \arg \min_{\mathbf{d} \in \mathcal{D}} R_n(\mathbf{d}(\mathbf{y})),$$

Can interpret this either by noting either

- that the Bayes risk is minimized when the inner integral is minimized for fixed \mathbf{y} , whatever the value of \mathbf{y} , the double integral is minimized, or*
- that we should directly choose the optimal decision conditional on the observed \mathbf{y} , and not average over all possible \mathbf{y} values.*

Decision theory concepts

For estimation, we need to choose the statistic T . We can reduce the problem to the minimization

$$\arg \min_{t \in \Theta} \int_{\Theta} L(t, \theta) \pi_n(\theta) d\theta$$

as whatever posterior the data lead to, we will minimize the double integral if we guarantee to minimize the inner integral.

Decision theory concepts

Thus, the decision that minimizes the Bayes risk minimizes *posterior expected loss* in making decision d , with expectation taken with respect to the posterior distribution $\pi_n(\theta)$.

Results for Different Loss Functions

(I) Under *squared-error loss*

$$L(t, \theta) = (t - \theta)^2$$

the *Bayes rule* for estimating θ is

$$\hat{d}_B(\mathbf{y}) = \hat{\theta}_{nB}(\mathbf{y}) = \mathbb{E}_{\pi_n}[\theta] = \int \theta \pi_n(\theta) d\theta$$

that is, the *posterior expectation*.

Results for Different Loss Functions

The *expected posterior loss* for any Bayes estimate t is

$$\int L(t, \theta) \pi_n(\theta) d\theta = \int (t - \theta)^2 \pi_n(\theta) d\theta$$

which needs to be minimized with respect to t .

Results for Different Loss Functions

We have

$$\begin{aligned}\frac{d}{dt} \left\{ \int (t - \theta)^2 \pi_n(\theta) d\theta \right\} &= \int \frac{d}{dt} \left\{ (t - \theta)^2 \right\} \pi_n(\theta) d\theta \\ &= \int 2(t - \theta) \pi_n(\theta) d\theta\end{aligned}$$

and equating this to zero gives

$$t = \int \theta \pi_n(\theta) d\theta = \mathbb{E}_{\pi_n} [\theta]$$

and hence the optimal $t = \hat{\theta}_{nB}$ is the posterior expectation as stated.

Results for Different Loss Functions

(II) Under *absolute error loss*

$$L(t, \theta) = |t - \theta|$$

the Bayes estimate for θ is the solution of

$$\int_{-\infty}^t \pi_n(\theta) d\theta = \frac{1}{2}$$

that is, it is the *posterior median*.

Results for Different Loss Functions

The expected posterior loss is

$$\int L(t, \theta) \pi_n(\theta) d\theta = \int |t - \theta| \pi_n(\theta) d\theta$$

which needs to be minimized with respect to t . We have that

$$\begin{aligned} & \int |t - \theta| \pi_n(\theta) d\theta \\ &= \int_{-\infty}^t (t - \theta) \pi_n(\theta) d\theta + \int_t^{\infty} (\theta - t) \pi_n(\theta) d\theta \end{aligned}$$

Results for Different Loss Functions

Differentiating with respect to t the first term yields

$$\begin{aligned} & \frac{d}{dt} \left\{ \int_{-\infty}^t (t - \theta) \pi_n(\theta) d\theta \right\} \\ &= \frac{d}{dt} \left\{ t \int_{-\infty}^t \pi_n(\theta) d\theta - \int_{-\infty}^t \theta \pi_n(\theta) d\theta \right\} \\ &= t\pi_n(t) + \int_{-\infty}^t \pi_n(\theta) d\theta - t\pi_n(t). \end{aligned}$$

Results for Different Loss Functions

Similarly

$$\frac{d}{dt} \left\{ \int_t^\infty (\theta - t) \pi_n(\theta) d\theta \right\} = -t\pi_n(t) - \int_t^\infty \pi_n(\theta) d\theta + t\pi_n(t)$$

Thus, equating the original derivative to zero yields

$$\int_{-\infty}^t \pi_n(\theta) d\theta - \int_t^\infty \pi_n(\theta) d\theta = 0$$

so that

$$\int_{-\infty}^t \pi_n(\theta) d\theta = \int_t^\infty \pi_n(\theta) d\theta = \frac{1}{2}$$

and hence the optimal $t = \hat{\theta}_{nB}$ is the *posterior median*.

Results for Different Loss Functions

(III) Under *zero-one loss*

$$L(t, \theta) = \begin{cases} 0 & t = \theta \\ 1 & t \neq \theta \end{cases}$$

the Bayes rule for estimating θ is

$$\hat{d}_B(\mathbf{y}) = \hat{\theta}_{nB}(\mathbf{y}) = \arg \max_{\theta \in \Theta} \pi_n(\theta)$$

that is, the *posterior mode*.

Results for Different Loss Functions

To see this, note that the expected posterior loss is

$$\int L(t, \theta) \pi_n(\theta) d\theta = \int_{\Theta \setminus t} \pi_n(\theta) d\theta$$

which needs to be minimized with respect to the choice of t . Consider the loss function

$$L_\delta(t, \theta) = \begin{cases} 0 & t \in (\theta - \delta, \theta + \delta) \\ 1 & t \notin (\theta - \delta, \theta + \delta) \end{cases}$$

for $\delta \geq 0$. That is, the loss is zero if $|t - \theta| < \delta$, and one otherwise.

Results for Different Loss Functions

The expected loss is therefore

$$\begin{aligned}\int L_\delta(t, \theta) \pi_n(\theta) d\theta &= \int_{\Theta \setminus (t-\delta, t+\delta)} \pi_n(\theta) d\theta \\ &= 1 - \Pr[\theta \in (t - \delta, t + \delta) | \mathbf{y}].\end{aligned}$$

Thus we need to choose t so that

$$\Pr[\theta \in (t - \delta, t + \delta) | \mathbf{y}]$$

is as large as possible, that is, we need to choose t as the centre of the highest posterior probability region of width 2δ . As $\delta \rightarrow 0$, this interval shrinks to be the posterior mode, as stated.

Likelihood Considerations

Likelihood Considerations

We have seen in the Bayesian calculation that the posterior distribution is highly dependent on the *likelihood* for its properties; on the log scale, we have in the iid case

$$\log \pi_n(\theta) = \sum_{i=1}^n \log f_Y(y_i; \theta) + \log \pi_0(\theta) + \text{constant}$$

and so as n grows, we expect the log-likelihood

$$\ell_n(\theta) = \sum_{i=1}^n \log f_Y(y_i; \theta)$$

to be the dominant term. Because of this it is useful to study the properties of the likelihood as n gets larger.

Asymptotic Theory of the Likelihood

Suppose that

- data $\mathbf{y} = (y_1, \dots, y_n)$ are realizations of iid random variables Y_1, \dots, Y_n drawn from distribution with pdf $f_0(y)$. We term this model the *true* model.
- we wish to represent the data using a parametric pdf $f_Y(y; \theta)$, where θ is d dimensional parameter. We term this model the *working* model.

Asymptotic Theory of the Likelihood

Typically, the analysis assumes that, for some θ_0 ,

$$f_0(y) \equiv f_Y(y; \theta_0)$$

that is, the parametric model is *correctly specified*.

However, if $f_0(y) \neq f_Y(y; \theta)$ for any θ , the model is *incorrectly specified*, and the theory needs to be reconsidered.

1. **Interpreting θ_0 in the working model:** We define the ‘true’ value of θ_0 as

$$\theta_0 = \arg \min_{\theta} KL(f_0, f_Y(\cdot; \theta)) \quad (3)$$

Note that

$$KL(f_0, f_Y(\cdot; \theta)) = \int \log f_0(y) f_0(y) dy - \int \log f_Y(y; \theta) f_0(y) dy$$

or equivalently, denoting $\log f_Y(y; \theta)$ by $\ell(y; \theta)$,

$$\theta_0 = \arg \max_{\theta} \mathbb{E}_{f_0} [\ell(Y; \theta)]. \quad (4)$$

2. **Maximum likelihood:** We maximize the sample-based expectation (or sample mean) to produce an estimator. Specifically, the estimator based on (4) will be

$$\hat{\theta}_n = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta).$$

This follows by the *weak law of large numbers*:

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta) \xrightarrow{P} \mathbb{E}_{f_0} [\ell(Y; \theta)] \quad (5)$$

as $n \rightarrow \infty$ for any fixed θ , if the expectation exists.

Asymptotic Theory of the Likelihood

We will assume that the log density $\ell(\mathbf{y}; \theta)$ is at least three times differentiable with respect to θ ; under this assumption, the estimate is defined as the solution to the *score equations*, the system of d equations given by

$$\frac{\partial}{\partial \theta} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{y}_i; \theta) \right\} = \mathbf{0}_d$$

or equivalently,

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \{ \ell(\mathbf{y}_i; \theta) \} = \frac{1}{n} \sum_{i=1}^n S(\mathbf{y}_i; \theta) = \mathbf{0}_d \quad (6)$$

say, where $S(\mathbf{y}; \theta) = \dot{\ell}(\mathbf{y}; \theta) = \partial \ell_1(\mathbf{y}; \theta) / \partial \theta$. Denote the solution of (6) by $\hat{\theta}_n \equiv \hat{\theta}_n(\mathbf{y})$.

3. **Taylor expansion:** We consider a Taylor expansion of the function $\ell(\mathbf{y}; \theta)$ with respect to θ around θ_0 .

$$\begin{aligned}\ell(\mathbf{y}; \theta) &= \ell(\mathbf{y}; \theta_0) + \dot{\ell}(\mathbf{y}; \theta_0)^\top (\theta - \theta_0) \\ &\quad + \frac{1}{2}(\theta - \theta_0)^\top \ddot{\ell}(\mathbf{y}; \theta_0)(\theta - \theta_0) + \mathcal{R}_3(\mathbf{y}; \theta^*)\end{aligned}\quad (7)$$

where

$$\ddot{\ell}(\mathbf{y}; \theta) = \frac{\partial^2 \ell(\mathbf{y}; \theta)}{\partial \theta \partial \theta^\top} \quad (d \times d).$$

and $\mathcal{R}_3(\mathbf{y}; \theta^*)$ is a remainder term, for some θ^* such that $\|\theta_0 - \theta^*\| \leq \|\theta_0 - \theta\|$.

Evaluating (7) for each of y_1, \dots, y_n and summing the result, we have

$$\begin{aligned}\ell_n(\theta) &= \ell_n(\theta_0) + \dot{\ell}_n(\theta_0)^\top (\theta - \theta_0) \\ &\quad + \frac{1}{2}(\theta - \theta_0)^\top \ddot{\ell}_n(\theta_0)(\theta - \theta_0) + \mathcal{R}_3.\end{aligned}\tag{8}$$

where $\mathcal{R}_3 \equiv \mathcal{R}_3(\mathbf{y}; \theta^*)$ for $\|\theta_0 - \theta^*\| \leq \|\theta_0 - \theta\|$.

At $\theta = \hat{\theta}_n$ and rearranging we have

$$\begin{aligned}\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) &= \dot{\ell}_n(\theta_0)^\top (\hat{\theta}_n - \theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\ell}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \mathcal{R}_3\end{aligned}\tag{9}$$

4. **Asymptotic behaviour:** Consider (9) written in terms of random variables, with $\hat{\theta}_n = \hat{\theta}_n(Y_{1:n})$:

$$\begin{aligned}\ell_n(\hat{\theta}_n) - \ell_n(\theta_0) &= \dot{\ell}_n(\theta_0)^\top (\hat{\theta}_n - \theta_0) \\ &\quad + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\ell}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \mathcal{R}_3\end{aligned}\tag{10}$$

Asymptotic Theory of the Likelihood

First consider for arbitrary θ , the quantity

$$\frac{1}{n} (\ell_n(\theta) - \ell_n(\theta_0)) = \frac{1}{n} \sum_{i=1}^n (\ell(Y_i; \theta) - \ell(Y_i; \theta_0)).$$

We may rewrite this expression with terms involving the true density f_0 that cancel :

$$\frac{1}{n} \sum_{i=1}^n (\ell(Y_i; \theta) - \ell_0(Y_i)) - \frac{1}{n} \sum_{i=1}^n (\ell(Y_i; \theta_0) - \ell_0(Y_i)) \quad (11)$$

where $\ell_0(y) = \log f_0(y)$.

Asymptotic Theory of the Likelihood

For any θ , as $n \rightarrow \infty$, we have by the weak law of large numbers that

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (\ell(Y_i; \theta) - \ell_0(Y_i)) &\xrightarrow{P} \mathbb{E}_{f_0} \left[\log \left(\frac{f_Y(Y; \theta)}{f_0(Y)} \right) \right] \\ &= -KL(f_0, f_Y(\cdot; \theta))\end{aligned}$$

as $Y_1, \dots, Y_n \sim f_0$.

Therefore

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta) - \frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta_0)$$

converges in probability to

$$KL(f_0, f_Y(\cdot; \theta_0)) - KL(f_0, f_Y(\cdot; \theta))$$

By definition of θ_0 via (3), $KL(f_0, f_Y(\theta))$ attains its minimum value at $\theta = \theta_0$, so

$$KL(f_0, f_Y(., \theta_0)) - KL(f_0, f_Y(., \theta)) \leq 0$$

and hence

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta) - \frac{1}{n} \sum_{i=1}^n \ell(Y_i; \theta_0)$$

converges in probability to a non-positive constant.

Therefore, we have that

$$\Pr_{f_0}[\ell_n(\theta_0) \geq \ell_n(\theta)] \longrightarrow 1 \quad (12)$$

as $n \longrightarrow \infty$. That is, with probability tending to 1, the log likelihood $\ell_n(\theta_0)$ is not less than $\ell_n(\theta)$ for any other θ .

Asymptotic Theory of the Likelihood

If we make an *identifiability* assumption, this statement may be strengthened: the model $f_Y(y; \theta)$ is *identifiable* if, for two parameter values $\theta^\dagger = \theta^\ddagger$,

$$f_Y(y; \theta^\dagger) = f_Y(y; \theta^\ddagger) \text{ for all } y \implies \theta^\dagger = \theta^\ddagger.$$

If the model is identifiable, then the “true” value θ_0 is uniquely defined, and we have

$$\Pr_{f_0}[\ell_n(\theta_0) > \ell_n(\theta)] \longrightarrow 1 \quad \theta \neq \theta_0. \quad (13)$$

Asymptotic Theory of the Likelihood

This theory holds for fixed θ_0 in the expression

$$\frac{1}{n} (\ell_n(\theta) - \ell_n(\theta_0))$$

However, we need to study $\ell_n(\hat{\theta}_n(Y_{1:n}))$, that is, where the parameter at which the log-likelihood is evaluated is itself a random variable, namely the estimator $\hat{\theta}_n(Y_{1:n})$.

Asymptotic Theory of the Likelihood

It can be shown that $\hat{\theta}_n(Y_{1:n}) \xrightarrow{p} \theta_0$ and $\hat{\theta}_n(Y_{1:n})$ is **consistent** for θ_0 , and by “continuous mapping” (as $\ell_n(\theta)$ is a continuous function in θ)

$$\left| \frac{1}{n} \left\{ \ell_n(\hat{\theta}_n(Y_{1:n})) - \ell_n(\theta_0) \right\} \right| \xrightarrow{p} 0$$

so that, from (5), as $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i; \hat{\theta}_n(Y_{1:n})) \xrightarrow{p} \mathbb{E}_{f_0} [\ell(Y; \theta_0)] \quad (14)$$

Asymptotic Theory of the Likelihood

5. **Asymptotic Normality:** For a continuous function such as $\dot{\ell}_n(\theta)$, with defined second derivative $\ddot{\ell}_n(\theta)$, it is guaranteed by the Mean Value Theorem that there exists an 'intermediate value'

$$\tilde{\theta} = c\hat{\theta}_n + (1 - c)\theta_0$$

for some c , $0 < c < 1$, such that

$$\dot{\ell}_n(\hat{\theta}_n) = \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\tilde{\theta})(\hat{\theta}_n - \theta_0)$$

The left hand side is zero as $\hat{\theta}_n$ is the mle.

Provided $\ddot{\ell}_n(\tilde{\theta})$ is non-singular, we may write after rescaling and rearrangement that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left\{ -\frac{1}{n} \ddot{\ell}_n(\tilde{\theta}) \right\}^{-1} \left\{ \sqrt{n} \left(\frac{1}{n} \dot{\ell}_n(\theta_0) \right) \right\} \quad (15)$$

Asymptotic Theory of the Likelihood

In its random variable form, second term on the right hand side of (15) is

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n S(Y_i; \theta_0) \right)$$

that is, a sample average quantity scaled by \sqrt{n} . But by definition of θ_0 ,

$$\mathbb{E}_{f_0}[S(Y; \theta_0)] = \int \dot{\ell}(y; \theta_0) f_0(y) dy = \mathbf{0}_d$$

as, by definition θ_0 minimizes $KL(f_0, f_Y(\cdot; \theta))$, and so must be a solution of this equation.

Therefore, by the Central Limit Theorem

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n S(Y_i; \theta_0) \right) \xrightarrow{d} \text{Normal}_d(\mathbf{0}_d, \mathcal{I}_{f_0}(\theta_0)) \quad (16)$$

where

$$\mathcal{I}_{f_0}(\theta_0) = \mathbb{E}_{f_0}[S(Y; \theta_0)S(Y; \theta_0)^\top] \equiv \text{Var}_{f_0}[S(Y; \theta_0)]$$

is a $(d \times d)$ quantity.

6. As $\hat{\theta}_n \xrightarrow{p} \theta_0$, we have that

$$-\frac{1}{n}\ddot{\ell}_n(\tilde{\theta}) \xrightarrow{\text{a.s.}} \mathcal{J}_{f_0}(\theta_0).$$

say.

Therefore from (15), we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left\{ -\frac{1}{n}\ddot{\ell}_n(\theta_0) \right\} \left\{ \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0) \right\} + o_p(1).$$

where $o_p(1)$ denotes a term that converges in probability to zero.

Asymptotic Theory of the Likelihood

The distribution of the second term is given by (16), and therefore have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{Normal}_d(\mathbf{0}_d, \Sigma(\theta_0))$$

where

$$\Sigma(\theta_0) = \{\mathcal{J}_{f_0}(\theta_0)\}^{-1} \mathcal{I}_{f_0}(\theta_0) \{\{\mathcal{J}_{f_0}(\theta_0)\}^{-1}\}^{\top}.$$

This follows as, for arbitrary \mathbf{Z} , if $\text{Var}[\mathbf{Z}] = \mathbf{V}$, then

$$\text{Var}[\mathbf{AZ}] = \mathbf{AVA}^{\top}$$

7. **Correct specification:** Under *correct specification*

$$f_0(y) \equiv f_Y(y; \theta_0),$$

and we have from earlier results that

$$\mathcal{I}_{\theta_0}(\theta_0) = \mathcal{J}_{\theta_0}(\theta_0)$$

and hence from the general result we deduce that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{Normal}_d(\mathbf{0}_d, \{\mathcal{I}_{\theta_0}(\theta_0)\}^{-1}).$$

Implications for Bayesian analysis

Using the same quadratic approximation for the likelihood at θ around $\hat{\theta}_n$ we have

$$\ell_n(\theta) \simeq \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)^\top (\hat{\theta}_n - \theta) + \frac{1}{2}(\hat{\theta}_n - \theta)^\top \ddot{\ell}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta)$$

but noting that $\dot{\ell}_n(\hat{\theta}_n) = 0$, we have that

$$\begin{aligned} \exp\{\ell_n(\theta)\} &\simeq \exp\{\ell_n(\hat{\theta}_n)\} \exp\left\{\frac{1}{2}(\hat{\theta}_n - \theta)^\top \ddot{\ell}_n(\hat{\theta}_n)(\hat{\theta}_n - \theta)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta - \hat{\theta}_n)^\top \{-\ddot{\ell}_n(\hat{\theta}_n)\}(\theta - \hat{\theta}_n)\right\}. \end{aligned}$$

Implications for Bayesian analysis

Thus, when the regularity conditions apply, the likelihood can be approximated by one arising from a Normal distribution

$$Normal_d \left(\hat{\theta}_n, \left\{ -\ddot{\ell}_n(\hat{\theta}_n) \right\}^{-1} \right).$$

This approximation can be used in a wide variety of models.

Modelling Extensions

Modelling Extensions

Beyond the iid case, Bayesian methods can be used for

- regression models (linear, non-linear, generalized linear);
- latent variable models;
- hierarchical models.

Regression models

We consider the infinite sequence $\{(X_n, Y_n), n = 1, 2, \dots\}$ such that for any $n \geq 1$

$$f_{X_1, \dots, X_n, Y_1, \dots, Y_n}(x_1, \dots, x_n, y_1, \dots, y_n)$$

is factorized

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) f_{Y_1, \dots, Y_n | X_1, \dots, X_n}(y_1, \dots, y_n | x_1, \dots, x_n)$$

where each term has a de Finetti representation.

$$\begin{aligned} f_{X_1, \dots, X_n}(\mathbf{x}_1, \dots, \mathbf{x}_n) \\ = \int \left\{ \prod_{i=1}^n f_X(\mathbf{x}_i; \phi) \right\} \pi_0(d\phi) \end{aligned}$$

$$\begin{aligned} f_{Y_1, \dots, Y_n | X_1, \dots, X_n}(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n) \\ = \int \left\{ \prod_{i=1}^n f_{Y|X}(y_i | \mathbf{x}_i; \theta) \right\} \pi_0(d\theta) \end{aligned}$$

Regression models

Inference for (ϕ, θ) is required:

- inference for ϕ via the *marginal model* for the X variables;
- inference for θ via the *conditional model* for Y given that $X = x$ was observed.

In the latter case, the fact that X is random is immaterial as we perform a conditional on x analysis.

When considering the statistical behaviour of Bayesian (or frequentist) procedures, we must remember that X and Y have *joint* structure.

Example: Prediction

To predict Y_{n+1} ,

$$\begin{aligned} f_{Y_{n+1}|X_{1:n}, Y_{1:n}}(y_{n+1} | x_{1:n}, y_{1:n}) \\ &= \int f_{X_{n+1}, Y_{n+1}|X_{1:n}, Y_{1:n}}(x_{n+1}, y_{n+1} | x_{1:n}, y_{1:n}) dx_{n+1} \\ &= \int f_{Y_{n+1}|X_{1:n}, X_{n+1}, Y_{1:n}}(y_{n+1} | x_{1:n}, x_{n+1}, y_{1:n}) \\ &\quad f_{X_{n+1}|X_{1:n}, Y_{1:n}}(x_{n+1} | x_{1:n}, y_{1:n}) dx_{n+1} \end{aligned}$$

Linear regression

Suppose that we have the linear regression model

$$Y_i = \mathbf{x}_i \beta + \varepsilon_i$$

where for $i = 1, \dots, n$

- Y_i is a scalar
- \mathbf{x}_i is $(1 \times d)$
- β is $(d \times 1)$
- $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$, independently.

Linear regression

This describes the model for the partially exchangeable

$$\epsilon_i = Y_i - \mathbf{x}_i\beta$$

random variables, conditional on the $\mathbf{X}_i = \mathbf{x}_i$.

There may or not be a need to model the distribution of the \mathbf{X}_i .

Linear regression

In vector form

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where \mathbf{Y} and ε are $(n \times 1)$, \mathbf{X} is $(n \times d)$.

We then have that in the conditional model

$$f_{Y_1, \dots, Y_n | X_1, \dots, X_n}(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n; \beta, \sigma^2) \equiv \text{Normal}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$$

where \mathbf{I}_n is the $(n \times n)$ identity matrix.

Linear regression

Therefore the likelihood is

$$\mathcal{L}_n(\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \right\}.$$

A conjugate prior in this setting can be factorized

$$\pi_0(\beta, \sigma^2) = \pi_0(\sigma^2)\pi_0(\beta|\sigma^2)$$

where

$$\pi_0(\sigma^2) \equiv \text{InvGamma}(a_0/2, b_0/2)$$

$$\pi_0(\beta|\sigma^2) \equiv \text{Normal}_d(\mathbf{m}_0, \sigma^2 \mathbf{M}_0)$$

where a_0, b_0, \mathbf{m}_0 and \mathbf{M}_0 are fixed constant *hyperparameters*.

$$\pi_0(\sigma^2) = \frac{(b_0/2)^{a_0/2}}{\Gamma(a_0/2)} \left(\frac{1}{\sigma^2} \right)^{a_0/2+1} \exp \left\{ -\frac{b_0}{2\sigma^2} \right\}$$

$$\begin{aligned} \pi_0(\beta|\sigma^2) = & \left(\frac{1}{2\pi\sigma^2} \right)^{d/2} \frac{1}{|\mathbf{M}_0|^{1/2}} \\ & \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \mathbf{m}_0)^\top \mathbf{M}_0^{-1} (\beta - \mathbf{m}_0) \right\} \end{aligned}$$

Linear regression

The joint posterior $\pi_n(\beta, \sigma^2)$ up to proportionality is

$$\mathcal{L}_n(\beta, \sigma^2)\pi_0(\beta, \sigma^2).$$

We examine the exponent as a quadratic form.

Linear regression

The expression

$$(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + (\beta - \mathbf{m}_0)^\top \mathbf{M}_0^{-1} (\beta - \mathbf{m}_0)$$

equates to

$$(\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1} (\beta - \mathbf{m}_n) + c_n$$

where we need to find expressions for \mathbf{m}_n , \mathbf{M}_n and c_n .

- Quadratic term:

$$\beta^\top \mathbf{M}_n^{-1} \beta = \beta^\top \mathbf{X}^\top \mathbf{X} \beta + \beta^\top \mathbf{M}_0^{-1} \beta$$

so therefore

$$\mathbf{M}_n^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{M}_0^{-1} \quad \therefore \quad \mathbf{M}_n = (\mathbf{X}^\top \mathbf{X} + \mathbf{M}_0^{-1})^{-1}$$

Linear regression

- Linear term:

$$\beta^\top \mathbf{M}_n^{-1} \mathbf{m}_n = \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{M}_0^{-1} \mathbf{m}_0$$

so therefore

$$\begin{aligned} \mathbf{m}_n &= \mathbf{M}_n (\mathbf{X}^\top \mathbf{y} + \mathbf{M}_0^{-1} \mathbf{m}_0) \\ &= (\mathbf{X}^\top \mathbf{X} + \mathbf{M}_0^{-1})^{-1} (\mathbf{X}^\top \mathbf{y} + \mathbf{M}_0^{-1} \mathbf{m}_0) \end{aligned}$$

Linear regression

- Constant term:

$$\mathbf{m}_n^\top \mathbf{M}_n^{-1} \mathbf{m}_n + c_n = \mathbf{y}^\top \mathbf{y} + \mathbf{m}_0^\top \mathbf{M}_0^{-1} \mathbf{m}_0$$

so therefore

$$c_n = \mathbf{y}^\top \mathbf{y} + \mathbf{m}_0^\top \mathbf{M}_0^{-1} \mathbf{m}_0 - \mathbf{m}_n^\top \mathbf{M}_n^{-1} \mathbf{m}_n$$

Linear regression

Therefore for the joint posterior up to proportionality is

$$\left(\frac{1}{\sigma^2}\right)^{\frac{(n+a_0+d)}{2}+1} \exp\left\{-\frac{(c_n + b_0)}{2\sigma^2}\right\} \\ \times \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1}(\beta - \mathbf{m}_n)\right\}$$

from which we can conclude directly that for the conditional posterior

$$\pi_n(\beta|\sigma^2) \equiv \text{Normal}_d(\mathbf{m}_n, \sigma^2 \mathbf{M}_n)$$

Linear regression

Integrating out β from the joint posterior, we obtain that up to proportionality

$$\pi_n(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{(n+a_0)}{2}+1} \exp\left\{-\frac{(c_n + b_0)}{2\sigma^2}\right\}$$

that is

$$\pi_n(\sigma^2) \equiv \text{InvGamma}(a_n/2, b_n/2)$$

where

$$a_n = n + a_0 \quad b_n = c_n + b_0$$

Linear regression

Finally, we can compute the marginal posterior for β . From the arguments above we have that the joint posterior takes the form

$$\pi_n(\beta|\sigma^2)\pi_n(\sigma^2)$$

which equates to

$$\begin{aligned} & \frac{(b_n/2)^{a_n/2}}{\Gamma(a_n/2)} \left(\frac{1}{\sigma^2}\right)^{a_n/2+1} \exp\left\{-\frac{b_n}{2\sigma^2}\right\} \\ & \times \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \frac{1}{|\mathbf{M}_n|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1}(\beta - \mathbf{m}_n)\right\} \end{aligned}$$

Linear regression

The constant term is

$$\frac{(b_n/2)^{a_n/2}}{\Gamma(a_n/2)} \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\mathbf{M}_n|^{1/2}}$$

and to marginalize we must compute

$$\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{a_n+d}{2}+1} \left\{ -\frac{1}{2\sigma^2} [b_n + (\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1} (\beta - \mathbf{m}_n)] \right\} d\sigma^2$$

Linear regression

The integrand is the kernel of an Inverse Gamma pdf so therefore we have that the integral equates to

$$\frac{\Gamma((a_n + d)/2)}{\left\{ \frac{1}{2} \left[b_n + (\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1} (\beta - \mathbf{m}_n) \right] \right\}^{\frac{a_n + d}{2}}}$$

Linear regression

Combining terms together, we have that

$$\pi_n(\beta) = \frac{b_n^{a_n/2}}{\Gamma(a_n/2)\pi^{d/2}} \frac{1}{|\mathbf{M}_n|^{1/2}} \frac{\Gamma((a_n + d)/2)}{\{b_n + (\beta - \mathbf{m}_n)^\top \mathbf{M}_n^{-1} (\beta - \mathbf{m}_n)\}^{\frac{a_n + d}{2}}}$$

which is a *multivariate Student-t distribution*.

Linear regression

Note that we may rewrite this form as

$$\pi_n(\beta) = \frac{\Gamma((a_n + d)/2)}{\Gamma(a_n/2) a_n^{d/2} \pi^{d/2}} \frac{1}{|\Sigma_n|^{1/2}} \frac{1}{\{1 + a_n^{-1}(\beta - \mathbf{m}_n)^\top \Sigma_n^{-1}(\beta - \mathbf{m}_n)\}^{\frac{a_n + d}{2}}}$$

which is the typical representation¹, where we have written

$$a_n \Sigma_n = b_n M_n$$

¹(https://en.wikipedia.org/wiki/Multivariate_t-distribution, as adopted in the R package mvnfast)

Note

We may express prior ignorance concerning β by considering $\mathbf{M}_0^{-1} \rightarrow \mathbf{0}$, in which case

$$\mathbf{m}_n \rightarrow (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\mathbf{M}_n \rightarrow (\mathbf{X}^\top \mathbf{X})^{-1}$$

yielding results equivalent to those of maximum likelihood.

This 'uniform' prior for β is in line with the earlier non-informative constructions.

Note

An alternative is the *g-prior*: for hyperparameter $\lambda > 0$

$$\mathbf{M}_0 = \lambda^{-1}(\mathbf{X}^\top \mathbf{X})^{-1}$$

in which case

$$\mathbf{M}_n = (1 + \lambda)^{-1}(\mathbf{X}^\top \mathbf{X})^{-1}$$

Linear regression

Note

If, for hyperparameter $\lambda > 0$

$$\mathbf{m}_0 = \mathbf{0}_d \quad \mathbf{M}_0 = \lambda \mathbf{I}_d$$

then

$$\mathbf{m}_n = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\mathbf{M}_n = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1}$$

yields the *ridge regression* procedure

Linear regression

Note

The log density is

$$\ell(\beta, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - \mathbf{x}\beta)^2 + \text{constant}$$

so therefore

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2} \mathbf{x}^\top (y - \mathbf{x}\beta)$$

$$\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial \beta \partial \beta^\top} = -\frac{1}{\sigma^2} \mathbf{x}^\top \mathbf{x}$$

Note

Also

$$\frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y - \mathbf{x}\beta)^2$$

$$\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(y - \mathbf{x}\beta)^2$$

and

$$\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \mathbf{x}^\top (y - \mathbf{x}\beta)$$

Note

Hence the (unit) Fisher information is

$$\mathcal{I}(\beta, \sigma^2) = \left| - \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{x}^\top \mathbf{x} & 0 \\ 0 & -\frac{1}{2\sigma^4} \end{bmatrix} \right| = \frac{1}{2} \left(\frac{1}{\sigma^2} \right)^{d+2} |\mathbf{x}^\top \mathbf{x}|$$

which implies that Jeffreys's prior for linear regression is

$$\pi_0(\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2} \right)^{d/2+1}$$

Linear regression

Note

This prior depends on dimension d . It is common instead to use the prior

$$\pi_0(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

as an invariant prior for linear regression.

Non-linear regression

- **Generalized Linear Models:** $f_{Y|X}(y|\mathbf{x}; \beta)$ follows an Exponential Family Model with

$$\mathbb{E}_{Y|X}[Y|\mathbf{X} = \mathbf{x}; \beta] = g^{-1}(\mathbf{x}\beta) \equiv \mu$$

$$\text{Var}_{Y|X}[Y|\mathbf{X} = \mathbf{x}; \beta] = V(\mu)$$

that is

$$g(\mu) = \mathbf{x}\beta$$

for some *link function*, g .

Non-linear regression

Example: Poisson regression

$$Y_i | \mathbf{X}_i = \mathbf{x}_i \sim \text{Poisson}(\mu_i)$$

$$\mathbb{E}_{Y|X}[Y_i | \mathbf{X}_i = \mathbf{x}_i; \beta] = \exp(\mathbf{x}_i \beta) \equiv \mu_i$$

$$\text{Var}_{Y|X}[Y | \mathbf{X}_i = \mathbf{x}_i; \beta] = \mu_i$$

so that

$$\mathcal{L}_n(\beta) = \prod_{i=1}^n \frac{\exp\{y_i \log \mu_i - \mu_i\}}{y_i!} = \prod_{i=1}^n \frac{\exp\{y_i \mathbf{x}_i \beta - \exp\{\mathbf{x}_i \beta\}\}}{y_i!}$$

Non-linear regression

Example: Poisson regression

$$\ell_n(\beta) = \sum_{i=1}^n (y_i \mathbf{x}_i \beta - \exp\{\mathbf{x}_i \beta\}) + \text{const.}$$

$$\dot{\ell}_n(\beta) = \sum_{i=1}^n (y_i \mathbf{x}_i^\top - \exp\{\mathbf{x}_i \beta\} \mathbf{x}_i^\top) = \sum_{i=1}^n \mathbf{x}_i^\top (y_i - \exp\{\mathbf{x}_i \beta\})$$

$$\ddot{\ell}_n(\beta) = - \sum_{i=1}^n \exp\{\mathbf{x}_i \beta\} \mathbf{x}_i^\top \mathbf{x}_i$$

that is, writing $\mathbf{D}(\mathbf{X}\beta) = \text{diag}(\exp\{\mathbf{x}_1 \beta\}, \dots, \exp\{\mathbf{x}_n \beta\})$.

$$\dot{\ell}_n(\beta) = \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}) \quad \ddot{\ell}_n(\beta) = -\mathbf{X}^\top \mathbf{D}(\mathbf{X}\beta) \mathbf{X}$$

Non-linear regression

Example: Binary regression

$$Y_i | \mathbf{X}_i = \mathbf{x}_i \sim \text{Bernoulli}(\mu_i)$$

$$\mathbb{E}_{Y|X}[Y_i | \mathbf{X} = \mathbf{x}_i; \beta] = \frac{\exp(\mathbf{x}_i \beta)}{1 + \exp(\mathbf{x}_i \beta)} \equiv \mu_i$$

$$\text{Var}_{Y|X}[Y_i | \mathbf{X}_i = \mathbf{x}_i; \beta] = \mu_i(1 - \mu_i)$$

Non-linear regression

Example: Binary regression

$$\begin{aligned}\mathcal{L}_n(\beta) &= \prod_{i=1}^n \exp \{y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)\} \\&= \prod_{i=1}^n \exp \left\{ y_i \log \left(\frac{\mu_i}{1 - \mu_i} \right) + \log(1 - \mu_i) \right\} \\&= \prod_{i=1}^n \exp \{y_i \mathbf{x}_i \beta - \log(1 + \exp\{\mathbf{x}_i \beta\})\}\end{aligned}$$

Non-linear regression

Example: Binary regression

$$\ell_n(\beta) = \sum_{i=1}^n (y_i \mathbf{x}_i \beta - \log(1 + \exp\{\mathbf{x}_i \beta\}))$$

$$\dot{\ell}_n(\beta) = \sum_{i=1}^n \mathbf{x}_i^\top \left(y_i - \frac{\exp\{\mathbf{x}_i \beta\}}{1 + \exp\{\mathbf{x}_i \beta\}} \right)$$

$$\ddot{\ell}_n(\beta) = - \sum_{i=1}^n \frac{\exp\{\mathbf{x}_i \beta\}}{(1 + \exp\{\mathbf{x}_i \beta\})^2} \mathbf{x}_i^\top \mathbf{x}_i = - \sum_{i=1}^n \mu_i (1 - \mu_i) \mathbf{x}_i^\top \mathbf{x}_i$$

that is, now writing $\mathbf{D}(\mathbf{X}\beta) = \text{diag}(\mu_1(1 - \mu_1), \dots, \mu_n(1 - \mu_n))$.

$$\dot{\ell}_n(\beta) = \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}) \quad \ddot{\ell}_n(\beta) = -\mathbf{X}^\top \mathbf{D}(\mathbf{X}\beta) \mathbf{X}$$

Non-linear regression

Using the quadratic approximation theory, we have that

$$\mathcal{L}_n(\beta) \simeq c_n(\hat{\beta}_n) \exp \left\{ -\frac{1}{2}(\beta - \hat{\beta}_n)^\top \Sigma_n^{-1}(\hat{\beta}_n)(\beta - \hat{\beta}_n) \right\}$$

where

$$\Sigma_n(\hat{\beta}_n) = \left(\mathbf{X}^\top \mathbf{D}(\mathbf{X}\hat{\beta}_n)\mathbf{X} \right)^{-1}$$

This approximate likelihood can be combined with a Normal prior on β .

Non-linear regression

Example: GLM

See knitr 3.

Non-linear regression

- **Non-linear regression:**

$$Y_i = g(\mathbf{x}_i; \theta) + \varepsilon_i$$

where $g(.,.)$ is some non-linear function of its arguments, and $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$.

$$\mathcal{L}_n(\theta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - g(\mathbf{x}_i; \theta))^2 \right\}.$$

Non-linear regression

Example: Exponential decay models

For $\theta = (\theta_0, \theta_1, \theta_3, \theta_4)^\top$ with $\theta_j > 0$ for $j = 1, 2, 3, 4$

$$g(\mathbf{x}_i; \theta) = \theta_1 e^{-\theta_2 x_i} + \theta_3 e^{-(\theta_2 + \theta_4) x_i}$$

where $x_i > 0$ is a scalar quantity.

- $\hat{\theta}_n$ found numerically;
- $\dot{\ell}(\theta, \sigma^2)$ and $\ddot{\ell}(\theta, \sigma^2)$ straightforward to compute;
- similar $Normal(\hat{\theta}_n, \Sigma_n(\hat{\theta}_n))$ approximation available.

Latent variable models

Latent (or *auxiliary*) variables can be introduced to simplify calculations in a model.

Suppose $f_Y(y; \theta)$ is intractable, but

$$f_Y(y; \theta) = \int f_{Y,Z}(y, z; \theta) \, dz$$

for some other variable Z , where the augmented joint distribution

$$f_{Y,Z}(y, z; \theta)$$

is tractable.

Latent variable models

Example: Mixture model

Suppose

$$f_Y(y; \theta) = (1 - \omega)f_0(y; \theta_0) + \omega f_1(y; \theta_1)$$

so that $\theta = (\omega, \theta_0, \theta_1)$, so that $0 < \omega < 1$. Then

$$f_Y(y; \theta) = \sum_{z=0}^1 f_{Y,Z}(y, z; \theta) = \sum_{z=0}^1 f_{Y|Z}(y|z; \theta) p_Z(z; \theta)$$

where

$$p_Z(z; \theta) = \Pr[Z = z] = \begin{cases} 1 - \omega & z = 0 \\ \omega & z = 1 \end{cases}$$

and

$$f_{Y|Z}(y|z; \theta) = f_z(y; \theta_z) \quad z = 0, 1.$$

Example: Mixture model

Then

$$f_{Y,Z}(y_i, z_i; \theta) = \omega^{z_i} (1 - \omega)^{1-z_i} f_0(y_i; \theta_0)^{1-z_i} f_1(y_i; \theta_1)^{z_i}$$

and the *sum* in the original pdf $f_Y(y; \theta)$ has become a *product*.

Example: Mixture model

Then

$$\prod_{i=1}^n f_Y(y_i; \theta) = \prod_{i=1}^n \{(1 - \omega)f_0(y_i; \theta_0) + \omega f_1(y_i; \theta_1)\}$$

which is not very tractable, but

$$\prod_{i=1}^n f_{Y,Z}(y_i, z_i; \theta) = \prod_{i=1}^n \prod_{z=0}^1 \{\omega_z f_z(y; \theta_z)\}^{\mathbb{1}_z(z_i)}$$

where $\omega_0 = (1 - \omega)$ and $\omega_1 = \omega$, which is more tractable.

Latent variable models

Example: see also

- data with *censoring*;
- *state space* models.

Hierarchical models

Hierarchical or *multi-level* models are built by ‘stacking’ levels of variables.

- *random effects* (or *mixed*) models;
- *multi-level* models
 - hospital/physician or school *league tables*;
 - *multi-arm* clinical studies;

Hierarchical models

Example: Multi-centre models

K centres, labelled $1, 2, \dots, K$.

- **STAGE 3:** For centre k , data Y_{k1}, \dots, Y_{kn_k} partially exchangeable, and conditionally independent given centre parameter θ_k . For each k

$$\prod_{i=1}^{n_k} f_k(y_{ki}; \theta_k).$$

- **STAGE 2:** Parameters $\theta_1, \dots, \theta_K$ exchangeable,

$$\prod_{k=1}^K \pi_0^{(2)}(\theta_k | \phi).$$

- **STAGE 1:** Prior on ϕ , $\pi_0^{(1)}(\phi)$.

Example: Multi-centre models

Data generating model:

- Pick $\phi \sim \pi_0^{(1)}(\phi)$
- Pick $\theta_1, \dots, \theta_K \sim \pi_0^{(2)}(\theta_k | \phi)$
- For each $k = 1, \dots, K$, pick

$$Y_{k1}, \dots, Y_{kn_k} \sim f_k(.; \theta_k)$$

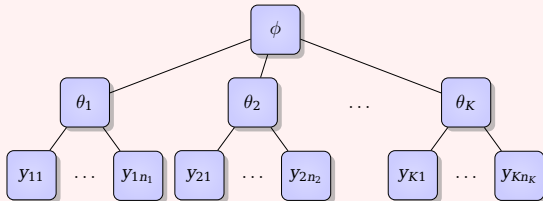
Hierarchical models

Example: Multi-centre models

STAGE 1 $\pi_0^{(1)}(\phi)$

STAGE 2 $\pi_2^{(2)}(\theta_k|\phi)$

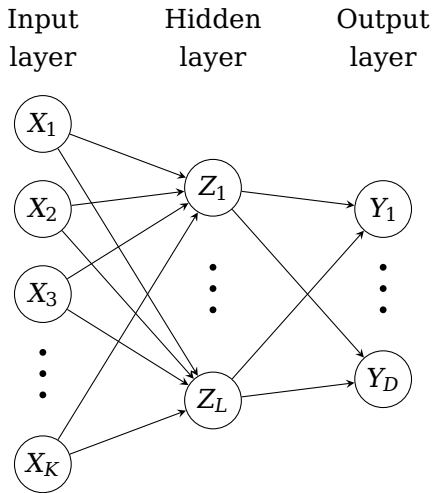
STAGE 3 $f_k(y_{ki};\theta_k)$



The posterior distribution $\pi_n(\phi, \theta_1, \dots, \theta_K)$ is given, up to proportionality, by

$$\pi_n(\phi, \theta_1, \dots, \theta_K) \propto \left\{ \prod_{k=1}^K \left\{ \prod_{i=1}^{n_k} f_k(y_{ki}; \theta_k) \right\} \pi_0^{(2)}(\theta_k | \phi) \right\} \pi_0^{(1)}(\phi)$$

Neural network models



Neural network models

Hidden Layer:

$$Z_l = g_{1l} \left(\sum_{k=1}^K w_{lk}^{(1)} X_k + b_l^{(1)}, \epsilon_l \right) \quad l = 1, \dots, L$$

with $\epsilon_1, \dots, \epsilon_L$ residual errors.

Output Layer:

$$Y_d = g_{2d} \left(\sum_{l=1}^L w_{dl}^{(2)} Z_l + b_d^{(2)}, \varepsilon_d \right) \quad d = 1, \dots, D$$

with $\varepsilon_1, \dots, \varepsilon_D$ residual errors.

Neural network models

- Data on X_1, \dots, X_K and Y_1, \dots, Y_D observed;
- Parameters are

$$\text{Weights} : w_{lk}^{(1)}, l = 1, \dots, L, k = 1, \dots, K$$
$$: w_{dl}^{(2)}, d = 1, \dots, D, l = 1, \dots, L$$

$$\text{Biases} : b_1^{(1)}, l = 1, \dots, L$$
$$: b_d^{(2)}, d = 1, \dots, D$$

- Link functions $g_{1l}(\cdot), l = 1, \dots, L$ and $g_{2d}(\cdot), d = 1, \dots, D$.

Neural network models

The *complete data* likelihood $\mathcal{L}_n(\mathbf{w}, \mathbf{b})$ is given, up to proportionality, by

$$\begin{aligned} \mathcal{L}_n(\mathbf{w}, \mathbf{b}) = & \prod_{i=1}^n \left\{ \prod_{l=1}^L \left\{ f_{Z_{li}|\mathbf{x}_i}(z_{li}|\mathbf{x}_i; \mathbf{w}^{(1)}, \mathbf{b}^{(1)}) \right\} \right. && (hidden) \\ & \times \left. \prod_{d=1}^D \left\{ f_{Y_{di}|\mathbf{z}_i}(y_{di}|\mathbf{z}_i; \mathbf{w}^{(2)}, \mathbf{b}^{(2)}) \right\} \right\} && (output) \end{aligned}$$

where the hidden variables

$$Z_{li}, l = 1, \dots, L, i = 1, \dots, n$$

are treated as auxiliary quantities.

Model Selection & Assessment

Model selection approaches

It is often necessary to consider model selection and evaluation approaches

- *in-sample* validity;
- *generalization*;

Assumptions

Consider the exchangeable, continuous case.

- For the *inference* model
 - $\theta \in \mathbb{R}^d$,
 - likelihood model $f_Y(y; \theta)$,
 - prior $\pi_0(\theta)$.
 - posterior $\pi_n(\theta)$.
- Suppose that the *data-generating* model is

$$f^*(y) \equiv f^*(y; \varphi)$$

with φ a fixed (but *unknown* to the *modeller*) value, so that exchangeability reduces to *independence*.

Predictive performance

The predictive distribution for the ‘next’ data point is

$$p_n(y) \equiv f_{Y_{n+1}|Y_1, \dots, Y_n}(y|y_1, \dots, y_n) = \int f_Y(y; \theta) \pi_n(\theta) d\theta$$

and is the usual Bayesian estimator of $f^*(y)$. It is used to assess the quality of a proposed model.

Predictive performance

If we consider instead

$$\tilde{p}_n(y) = f_{Y_{n+1}|Y_1, \dots, Y_n}(y|Y_1, \dots, Y_n)$$

then the predictive distribution itself is a *random function*, as it is a function of rvs Y_1, \dots, Y_n , not the data y_1, \dots, y_n .

We may similarly consider the *random* posterior $\tilde{\pi}_n(\theta)$, a function of θ that is random because its inputs are Y_1, \dots, Y_n instead of y_1, \dots, y_n .

Predictive performance

The KL divergence between $f^*(y)$ and $p_n(y)$ is

$$\begin{aligned} KL(f^*, p_n) &= \int \log \left(\frac{f^*(y)}{p_n(y)} \right) f^*(y) \, dy \\ &= \int \log(f^*(y)) f^*(y) \, dy - \int \log(p_n(y)) f^*(y) \, dy. \end{aligned} \quad (\diamond)$$

The first term in (\diamond) is a constant which does not depend on the inference model.

A random variable version $KL(f^*, \tilde{p}_n)$ can also be considered.

Predictive performance

- **Training loss:** The *training loss*, T_n , is a measure that approximates the KL divergence based on the sample

$$T_n \equiv T(Y_1, \dots, Y_n) = -\frac{1}{n} \sum_{i=1}^n \log \tilde{p}_n(Y_i)$$

which can be regarded as a sample-based estimator of the second term in (\diamond) , with the data drawn independently from f^* .

In this form, T_n is random variable as it depends on \tilde{p}_n .

- **Generalization loss:** The *generalization loss*, G_n , is the second term in (\diamond):

$$G_n \equiv G(Y_1, \dots, Y_n) = - \int \log \tilde{p}_n(y) f^*(y) \, dy.$$

This can only be computed precisely if $f^*(y)$ is known. However, we can interpret G_n as a measure of proximity of the predictive model to the data-generating distribution.

Note

The first term in (\diamond) is often denoted $-S$

$$S = - \int \log(f^*(y)) f^*(y) dy$$

and is termed the *entropy* of f^* . The quantity

$$G_n - S$$

is termed the *generalization error*: note that $G_n \geq S$ as the KL divergence is non-negative.

Predictive performance

- **Cross-validation loss:** The *cross-validation* loss, C_n , is defined by

$$C_n = -\frac{1}{n} \sum_{i=1}^n \log \tilde{p}_n^{(-i)}(Y_i)$$

where $\tilde{p}_n^{(-i)}(y)$ is the posterior predictive distribution derived from the random variables

$$Y_{1:n}^{(-i)} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$$

that is, the original collection with Y_i removed.

Predictive performance

Taking expectations of G_n and C_n with respect to the joint pdf of Y_1, \dots, Y_n , which by independence reduces to

$$\prod_{i=1}^n f^*(y_i)$$

we can establish connections between the losses.

Predictive performance

Provided all expectations are finite

$$\begin{aligned}\mathbb{E}[C_n] &= \mathbb{E}_{Y_1, \dots, Y_n} \left[-\frac{1}{n} \sum_{i=1}^n \log \tilde{p}_n^{(-i)}(Y_i) \right] \\ &= -\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{1:n}^{(-i)}} \left[\mathbb{E}_{Y_i} \left[\log \tilde{p}_n^{(-i)}(Y_i) \right] \right] \\ &= -\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{1:n}^{(-i)}} \left[\int \log \tilde{p}_n^{(-i)}(y) f^*(y) dy \right]\end{aligned}$$

where the second line follows using iterated expectation, and as Y_1, \dots, Y_n are independently drawn from f^* .

Predictive performance

But for $i = 1, 2, \dots, n$ the scalar quantities

$$\int \log \tilde{p}_n^{(-i)}(y) f^*(y) dy$$

are *identically distributed* random variables; recall that

$$\tilde{p}_n^{(-i)}(y)$$

is a random variable for each y as it is a function of

$$Y_{1:n}^{(-i)} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n).$$

Therefore

$$-\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y_{1:n}^{(-i)}} \left[\int \log \tilde{p}_n^{(-i)}(y) f^*(y) dy \right]$$

is equal to

$$\mathbb{E}_{Y_{1:n}^{(-1)}} \left[- \int \log \tilde{p}_n^{(-1)}(y) f^*(y) dy \right] \equiv \mathbb{E}[G_{n-1}]$$

again as Y_1, \dots, Y_n are iid from f^* .

Predictive performance

Note also that

$$\begin{aligned} p_n^{(-i)}(y) &= \int f_Y(y; \theta) \pi_n^{(-i)}(\theta) d\theta \\ &= \int f_Y(y; \theta) \frac{\prod_{j \neq i} f_Y(y_j; \theta) \pi_0(\theta)}{\int \prod_{j \neq i} f_Y(y_j; t) \pi_0(t) dt} d\theta \end{aligned}$$

so therefore

$$\tilde{p}_n^{(-i)}(Y_i) = \frac{\int f_Y(Y_i; \theta) \prod_{j \neq i} f_Y(Y_j; \theta) \pi_0(\theta) d\theta}{\int \prod_{j \neq i} f_Y(Y_j; t) \pi_0(t) dt}$$

Predictive performance

Numerator:

$$\int f_Y(Y_i; \theta) \prod_{j \neq i} f_Y(Y_j; \theta) \pi_0(\theta) d\theta = \int \prod_{j=1}^n f_Y(Y_j; \theta) \pi_0(\theta) d\theta$$

Denominator:

$$\int \prod_{j \neq i} f_Y(Y_j; t) \pi_0(t) dt = \int \frac{1}{f_Y(Y_i; t)} \prod_{j=1}^n f_Y(Y_j; t) \pi_0(t) dt$$

Predictive performance

Therefore

$$\begin{aligned}C_n &= -\frac{1}{n} \sum_{i=1}^n \log \tilde{p}_n^{(-i)}(Y_i) \\&= \frac{1}{n} \sum_{i=1}^n \log \frac{\int \frac{1}{f_Y(Y_i; t)} \prod_{j=1}^n f_Y(Y_j; t) \pi_0(t) dt}{\int \prod_{j=1}^n f_Y(Y_j; \theta) \pi_0(\theta) d\theta} \\&= \frac{1}{n} \sum_{i=1}^n \log \int \frac{1}{f_Y(Y_i; t)} \frac{\prod_{j=1}^n f_Y(Y_j; t) \pi_0(t)}{\int \prod_{j=1}^n f_Y(Y_j; \theta) \pi_0(\theta) d\theta} dt\end{aligned}$$

Predictive performance

But t and θ are merely dummy integrating variables, so we may exchange them and write

$$\begin{aligned} C_n &= \frac{1}{n} \sum_{i=1}^n \log \int \frac{1}{f_Y(Y_i; \theta)} \frac{\prod_{j=1}^n f_Y(Y_j; \theta) \pi_0(\theta)}{\int \prod_{j=1}^n f_Y(Y_j; t) \pi_0(t) dt} d\theta \\ &= \frac{1}{n} \sum_{i=1}^n \log \mathbb{E}_{\tilde{\pi}_n} \left[\frac{1}{f_Y(Y_i; \theta)} \right] \end{aligned}$$

as the *term in red* is merely the random variable version of the posterior $\tilde{\pi}_n(\theta)$.

Predictive performance

This identity may be useful as it gives an expression for computing the numerical value of C_n which does not depend on the *leave-one-out* posterior distributions:

- the original formula requires n separate posterior calculations of the quantities $p_n^{(-i)}(y)$;
- the new formula requires only the computation of $\pi_n(\theta)$, the full posterior;
- the new formula does require the computation of

$$\mathbb{E}_{\pi_n} \left[\frac{1}{f_Y(y_i; \theta)} \right]$$

for $i = 1, \dots, n$.

Predictive performance

- **WAIC:** The *widely applicable information criterion* (or WAIC), W_n , is defined by

$$W_n = T_n + \frac{1}{n} \sum_{i=1}^n \text{Var}_{\tilde{\pi}_n} [\log f_Y(Y_i; \theta)]$$

where, recall, T_n is the training loss

$$T_n = -\frac{1}{n} \sum_{i=1}^n \log \tilde{p}_n(Y_i)$$

Note

It can be shown that

$$W_n = C_n + O_p\left(\frac{1}{n^2}\right)$$

and so W_n provides a tractable approximation strategy.

- **Marginal likelihood (or prior predictive):** The normalizing constant that appears in the denominator of the (random) posterior $\tilde{\pi}_n(\theta)$ is

$$Z_n \equiv Z(Y_1, \dots, Y_n) = \int \prod_{i=1}^n f_Y(Y_i; \theta) \pi_0(\theta) d\theta.$$

and, by de Finetti, this can be interpreted as the value of the (random) joint pdf

$$f_{Y_{1:n}}(Y_{1:n}) \equiv f_{Y_1, \dots, Y_n}(Y_1, \dots, Y_n).$$

Predictive performance

The quantity Z_n is termed the

- *marginal likelihood*,
- *prior predictive* distribution.

In this form, $Z_n = Z(Y_1, \dots, Y_n)$ is a random variable:

$$z_n = Z(y_1, \dots, y_n)$$

can also be computed.

Note that

$$KL(f_{Y_{1:n}}^*, f_{Y_{1:n}}) = \int \log \left(\frac{f_{Y_{1:n}}^*(y_{1:n})}{f_{Y_{1:n}}(y_{1:n})} \right) f_{Y_{1:n}}^*(y_{1:n}) \, dy_{1:n}$$

measures the divergence between the data-generating joint pdf

$$f_{Y_{1:n}}^*(y_{1:n}) = \prod_{i=1}^n f_Y^*(y_i)$$

and the modelled joint pdf $f_{Y_{1:n}}(y_{1:n})$.

Thus

$$\begin{aligned} KL(f_{Y_{1:n}}^*, f_{Y_{1:n}}) &= \int \log f_{Y_{1:n}}^*(y_{1:n}) f_{Y_{1:n}}^*(y_{1:n}) \, dy_{1:n} \\ &\quad - \int \log f_{Y_{1:n}}(y_{1:n}) f_{Y_{1:n}}^*(y_{1:n}) \, dy_{1:n} \end{aligned}$$

for which the term being subtracted is

$$\mathbb{E}_{f_{Y_{1:n}}^*} [\log f_{Y_{1:n}}(Y_{1:n})] = \mathbb{E}[\log Z_n].$$

Predictive performance

The random variable

$$F_n = -\log Z_n$$

that is, minus the log marginal likelihood, is sometimes termed the *free energy*.

Predictive performance

We have that

$$\begin{aligned} p_n(y_{n+1}) &= \int f_Y(y_{n+1}; \theta) \frac{\prod_{i=1}^n f_Y(y_i; \theta) \pi_0(\theta)}{\int \prod_{i=1}^n f_Y(y_i; t) \pi_0(t) dt} d\theta \\ &= \frac{\int \prod_{i=1}^{n+1} f_Y(y_i; \theta) \pi_0(\theta) d\theta}{\int \prod_{i=1}^n f_Y(y_i; t) \pi_0(t) dt} \\ &= \frac{Z_{n+1}}{Z_n} \end{aligned}$$

Predictive performance

Therefore

$$\log \tilde{p}_n(Y_{n+1}) = \log Z_{n+1} - \log Z_n = F_n - F_{n+1}.$$

Predictive performance

Note

Note that by direct calculation, we have

$$\mathbb{E}[G_n] = \mathbb{E}[F_{n+1}] - \mathbb{E}[F_n]$$

or equivalently

$$\mathbb{E}[F_n] = \mathbb{E}[F_1] + \sum_{i=1}^{n-1} \mathbb{E}[G_i]$$

Predictive performance

The quantities

- T_n
- G_n
- C_n
- W_n
- F_n

can all be used for model evaluation and comparison.

Other Interpretations of Bayes

Variational formulations

Objective: choose a distribution $p_{\mathbf{Y}}(\mathbf{y})$ for the observable collection $\mathbf{y} = (y_1, \dots, y_n)$ using the principle of *entropy minimization*.

Choose $p_{\mathbf{Y}}$ by minimizing

$$- \int \log p_{\mathbf{Y}}(\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}$$

subject to the constraint that

$$p_{\mathbf{Y}}(\mathbf{y}) = \int p_{\mathbf{Y},\theta}(\mathbf{y}, \theta) d\theta$$

for parameter θ .

Variational formulations

If $\pi(\cdot)$ is an arbitrary density in θ indexed by \mathbf{y} , we have

$$\begin{aligned}\log p_{\mathbf{Y}}(\mathbf{y}) &= \log \int p_{\mathbf{Y},\theta}(\mathbf{y}, \theta) \, d\theta \\ &= \log \int \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\pi(\theta)} \pi(\theta) \, d\theta \\ &\geq \int \log \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\pi(\theta)} \pi(\theta) \, d\theta\end{aligned}$$

by Jensen's Inequality.

Variational formulations

Therefore

$$\begin{aligned} - \int \log p_{\mathbf{Y}}(\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) \, d\mathbf{y} &= - \int \log \left\{ \int p_{\mathbf{Y},\theta}(\mathbf{y}, \theta) \, d\theta \right\} \, d\mathbf{y} \\ &= - \int \log \left\{ \int \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\pi(\theta)} \pi(\theta) \, d\theta \right\} \, d\mathbf{y} \\ &= - \int \log \mathbb{E}_{\pi} \left[\int \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\pi(\theta)} \right] \, d\mathbf{y} \\ &\geq \int \mathbb{E}_{\pi} \left[\log \frac{\pi(\theta)}{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)} \right] \, d\mathbf{y} \end{aligned}$$

Variational formulations

The bound is therefore achieved if

$$\pi(\theta) \propto p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)$$

so that

$$\pi(\theta) = \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\int p_{\mathbf{Y},\theta}(\mathbf{y}, t) dt} \equiv \pi_n(\theta|\mathbf{y}).$$

This holds for any \mathbf{y} , and shows that the entropy minimizing choice for $p_{\mathbf{Y}}(\mathbf{y})$ requires the representation

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{p_{\mathbf{Y},\theta}(\mathbf{y}, \theta)}{\pi_n(\theta|\mathbf{y})}$$

Variational formulations

Recall that the Kullback-Leibler divergence between densities p and q is

$$KL(p, q) = \int \log \frac{p(y)}{q(y)} p(y) dy.$$

This is a non-negative quantity, and is finite provided $q(y) > 0$ whenever $p(y) > 0$.

Variational formulations

Lemma

Consider two distributions p and q with common support \mathcal{Y} , and let T be a generic scalar function of Y . Then the KL divergence admits the following dual representation

$$KL(p, q) = \sup_{T: \mathcal{X} \rightarrow \mathbb{R}} \{ \mathbb{E}_p[T] - \log \mathbb{E}_q[\exp\{T\}] \}$$

provided the two expectations are finite.

We can apply the Lemma to deduce a **variational inequality** relevant to Bayes theorem.

Variational formulations

For fixed \mathbf{y} , define function $T_{\mathbf{y}}(\theta) = g(\mathbf{y}, \theta)$. We have for all putative ‘posterior’ distributions $\pi(\theta)$ that

$$KL(\pi, \pi_0) \geq \mathbb{E}_{\pi}[g(\mathbf{y}, \theta)] - \log \mathbb{E}_{\pi_0}[\exp\{g(\mathbf{y}, \theta)\}].$$

Therefore

$$\begin{aligned} -\log \int \exp\{g(\mathbf{y}, \theta)\} \pi_0(\theta) \, d\theta &\leq \int -g(\mathbf{y}, \theta) \pi(\theta) \, d\theta \\ &\quad + \int \log \left(\frac{\pi(\theta)}{\pi_0(\theta)} \right) \pi(\theta) \, d\theta \end{aligned}$$

We aim to find the π that is ‘close’ to π_0 in light of the data.

Variational formulations

The bound is achieved if

$$g(\mathbf{y}, \theta) = \log \frac{\pi(\theta)}{\pi_0(\theta)} + C(\mathbf{y})$$

so that

$$\pi(\theta) = \frac{\exp\{g(\mathbf{y}, \theta)\} \pi_0(\theta)}{c(\mathbf{y})}.$$

Thus it is evident that

$$c(\mathbf{y}) = \exp\{-C(\mathbf{y})\} = \int \exp\{g(\mathbf{y}, \theta)\} \pi_0(\theta) \, d\theta.$$

Variational formulations

For this choice of π , the bound is achieved and

$$KL(\pi, \pi_0) = \mathbb{E}_{\pi}[g(\mathbf{y}, \theta)] - \log \mathbb{E}_{\pi_0}[\exp\{g(\mathbf{y}, \theta)\}]$$

The Bayesian posterior follows if g is the log likelihood.

Consider an *Information Processing Rule* (IPR) for which

- we input the prior π_0 and the likelihood \mathcal{L}_y ,
- the rule returns the posterior π and marginal $p_Y \equiv p_0$.

Required to find an IPR where no information is lost or gained during the functional conversion.

Information processing

Information content of the various components:

$$\text{Information in } \mathcal{L}_{\mathbf{y}}(\theta) : \int \log \mathcal{L}_{\mathbf{y}}(\theta) \pi(\theta) d\theta \quad (\text{I1})$$

$$\text{Information in } \pi_0(\theta) : \int \log \pi_0(\theta) \pi(\theta) d\theta \quad (\text{I2})$$

$$\text{Information in } \pi(\theta) : \int \log \pi(\theta) \pi(\theta) d\theta \quad (\text{O1})$$

$$\text{Information in } p_0(\mathbf{y}) : \int \log p_0(\mathbf{y}) \pi(\theta) d\theta = \log p_0(\mathbf{y}) \quad (\text{O2})$$

We wish to *choose* $\pi(\theta)$ to ensure that

$$\text{I1} + \text{I2} = \text{O1} + \text{O2}.$$

The *information discrepancy* for a putative posterior π is

$$\int \log \pi(\theta) \pi(\theta) \, d\theta + \log p_0(\mathbf{y}) - \int [\log \mathcal{L}_{\mathbf{y}}(\theta) + \log \pi_0(\theta)] \pi(\theta) \, d\theta$$

which can be re-written

$$D(\pi) = \int \log \left\{ \frac{\pi(\theta) p_0(\mathbf{y})}{\pi_0(\theta) \mathcal{L}_{\mathbf{y}}(\theta)} \right\} \pi(\theta) \, d\theta.$$

By properties of the Kullback-Leibler divergence, this is minimized in π (and becomes zero) when

$$\pi(\theta) = \frac{\pi_0(\theta)\mathcal{L}_{\mathbf{y}}(\theta)}{p_0(\mathbf{y})}$$

that is, π is the Bayesian posterior.

However

- (I1) is not readily interpreted as an information quantity concerning θ as $\mathcal{L}_{\mathbf{y}}(\theta)$ is not itself a probability distribution in θ .
- Alternative justification: our task is to report a ‘decision’ concerning a feature of interest in light of data \mathbf{y} based on a *utility function*, u that takes as arguments a (putative posterior) probability distribution $\pi(\cdot)$ and a parameter value θ and returns a non-negative real value.

- (i) *Proper*: for any other π' , the expected utility wrt π ,

$$\int u(\pi', \theta) \pi(\theta) d\theta$$

is uniquely maximized when $\pi' \equiv \pi$.

- (ii) *Local*: the utility function satisfies

$$u(\pi, \theta) = u(\pi(\theta), \theta)$$

that is, that the utility value depends only on the utility function evaluated at θ .

Utility

These conditions imply that u takes the form

$$u(\pi, \theta) = A \log \pi(\theta) + B(\theta)$$

Under this *log* utility, the optimal π turns out to be $\pi_n(\theta|\mathbf{y})$.

Choose the rule $\pi_{\mathbf{y}}(\theta)$ (a distribution indexed by arbitrary \mathbf{y})

$$\begin{aligned}\pi_{\mathbf{y}}(\theta) &= \arg \max_{\pi} \iint u(\pi, \theta) f_{\mathbf{Y}}(\mathbf{y}; \theta) \pi_0(\theta) d\theta d\mathbf{y} \\ &= \arg \max_{\pi} \iint \log \pi(\theta) f_{\mathbf{Y}}(\mathbf{y}; \theta) \pi_0(\theta) d\theta d\mathbf{y} \\ &= \arg \min_{\pi} \iint \log \frac{\pi_0(\theta) f_{\mathbf{Y}}(\mathbf{y}; \theta)}{\pi(\theta)} f_{\mathbf{Y}}(\mathbf{y}; \theta) \pi_0(\theta) d\theta d\mathbf{y}\end{aligned}$$

which, by properties of the Kullback-Leibler divergence, implies that

$$\pi_{\mathbf{y}}(\theta) \propto f_{\mathbf{Y}}(\mathbf{y}; \theta) \pi_0(\theta).$$

Non-probabilistic methods

Recently, there has been increasing focus on *non-probabilistic* updating rules that do not model

$$f(y; \theta)$$

directly, but recognize the issue of mis-specification.

Task: report a ‘posterior’ distribution $\pi_n^\dagger(\theta)$ that

- updates prior to posterior in light of data
- respects common assumptions wrt iid data etc.

Non-probabilistic methods

The *'Gibbs' posterior* $\pi_n^\dagger(\theta)$ is computed as

$$\pi_n^\dagger(\theta) = \arg \min_{\pi \in \mathcal{M}} \left\{ \int \ell(y_{1:n}, \theta) \pi(d\theta) + \lambda \int \log \left(\frac{d\pi(\theta)}{d\pi_0^\dagger(\theta)} \right) \pi(d\theta) \right\}$$

- the first term is the expected “post-data” loss

$$\ell(y_{1:n}, \theta)$$

- the second term is the KL divergence between prior π_0^\dagger and candidate Gibbs posterior π .

Constant λ makes the two losses commensurate.

Non-probabilistic methods

If $\eta = 1/\lambda$, then

$$\pi_n^\dagger(\theta) = \frac{\exp\{-\eta\ell(y_{1:n}, \theta)\}\pi_0^\dagger(\theta)}{\int \exp\{-\eta\ell(y_{1:n}, t)\}\pi_0^\dagger(t) dt}.$$

- The only legitimate form for iid data is

$$\ell(y_{1:n}, \theta) = \sum_{i=1}^n \ell(y_i, \theta).$$

- $\pi_n^\dagger(\theta)$ and $\pi_n(\theta)$ coincide if

$$\eta\ell(y, \theta) = -\log f(y; \theta)$$

for a parametric model.

Non-probabilistic methods

This formulation is potentially attractive as it allows us to target θ directly:

- even if data generating F is ‘big and complicated’, θ may be low-dimensional,
- this method avoids having to develop full model.

aka *Generalized Bayesian inference*.

Non-probabilistic methods

Power posterior: if $\ell(\mathbf{y}, \theta) = -\log f(\mathbf{y}; \theta)$, then

$$\pi_n^\dagger(\theta) \propto \prod_{i=1}^n \{f(\mathbf{y}_i; \theta)\}^\eta \pi_0(\theta)$$

where η allows for mis-specification of the selected model.

NB: $\{f(\mathbf{y}; \theta)\}^\eta$ is not a pdf: normalizing constant

$$\int \{f(\mathbf{y}; \theta)\}^\eta d\mathbf{x}$$

may depend on θ .

Also need to choose η .

Non-probabilistic methods

The *Gibbs* posterior is

$$\pi_n^\dagger(\theta) = \frac{\exp \left\{ -\eta \sum_{i=1}^n \ell(\mathbf{y}_i, \theta) \right\} \pi_0^\dagger(\theta)}{\int \exp \left\{ -\eta \sum_{i=1}^n \ell(\mathbf{y}_i, t) \right\} \pi_0^\dagger(t) dt}$$

whereas the *standard* Bayesian calculation yields

$$\pi_n(\theta) = \frac{\exp \left\{ \sum_{i=1}^n \log f(\mathbf{y}_i; \theta) \right\} \pi_0(\theta)}{\int \exp \left\{ \sum_{i=1}^n \log f(\mathbf{y}_i; t) \right\} \pi_0(t) dt}$$

Non-probabilistic methods

- Y_1, \dots, Y_n, \dots are iid from F_0 ,
- F_0 is unknown
- problem admits a de Finetti representation.

Non-probabilistic methods

Suppose $\theta = (\psi, \zeta)$. Marginally

$$\pi_n(\psi) = \frac{\int \exp \left\{ \sum_{i=1}^n \log f(y_i; \psi, \zeta) \right\} \pi_0(\psi, \zeta) d\zeta}{\int \exp \left\{ \sum_{i=1}^n \log f(y_i; t, s) \right\} \pi_0(t, s) dt ds}$$

Comparing $\pi_n(\psi)$ and $\pi_n^\dagger(\psi)$ we might interpret $\pi_n^\dagger(\psi)$ as an attempt to represent the true marginal model

$$\exp \left\{ -\eta \sum_{i=1}^n \ell(\mathbf{y}_i, \psi) \right\} \approx \int \exp \left\{ \sum_{i=1}^n \log f(\mathbf{y}_i; \psi, \zeta) \right\} \pi_0(\zeta|\psi) \, d\zeta$$

Issues with this formulation

- Does not give a coherent prediction mechanism;
 - No assumption that

$$c(\theta) = \int \exp\{-\eta \ell(y, \theta)\} dx < \infty$$

- Even if $c(\theta)$ is finite, in general it is a function of θ .
- Does not give a ready interpretation of θ in general.

Targeting parameters via loss functions

Define

$$\theta_0 = \arg \min_{t \in \mathcal{T}} \int \ell(y, t) dF_0(y).$$

where

- θ_0 is a *loss function-defined parameter*
- θ_0 does *not* characterize F_0 in general.

Alternatively: may define θ_0 via

$$\int u(y, \theta_0) dF_0(y) = 0$$

for some suitably defined $u(\cdot, \cdot)$.

Targeting parameters via loss functions

Example:

- $Y \sim \text{Poisson}(\lambda_0)$
- $\ell(y, \theta) = \sqrt{y}(y - \theta)^2$
-

$$\theta_0 = \frac{\sum_{y=0}^{\infty} \sqrt[3]{y} f(y; \lambda_0)}{\sum_{y=0}^{\infty} \sqrt{y} f(y; \lambda_0)}$$

‘Easy’ to compute a posterior $\pi_n(\theta)$ from $\pi_n(\lambda)$.

Survival Example

In a survival problem, we might be interested in a *median* lifetime defined via the loss

$$\ell(y, \theta) = |y - \theta|$$

that we learn about having fitted

- an *Exponential*(λ) model
- a *Weibull*(α, λ) model
- a *LogNormal*(μ, σ^2) model
- a non-parametric model

Survival Example

All of these models yield the posterior for F that we can convert into a posterior for θ

$$\theta = F^{-1}(1/2)$$

by transformation.

Bayesian inference for targeted parameters

In the parametric case: $F_0(y) \equiv F(y; \vartheta_0)$ say.

- **Prior:**

- (i) Sample $\vartheta^{(b)} \sim \pi_0$
- (ii) Convert this to

$$\theta^{(b)} = \arg \min_{t \in \mathcal{T}} \int \ell(y, t) dF(y; \vartheta^{(b)})$$

via a deterministic transformation.

- **Posterior:** $\pi_n(\vartheta)$ defines $\pi_n(\theta)$ in a similar fashion.

Bayesian inference for targeted parameters

In the non-parametric case: $F(\cdot)$ is the parameter

- **Prior:**

- (i) Sample $F^{(b)} \sim \pi_0^F$
- (ii) Convert this to

$$\theta^{(b)} = \arg \min_{t \in \mathcal{T}} \int \ell(y, t) dF^{(b)}(y)$$

via a deterministic transformation.

- **Posterior:** $\pi_n^F(F)$ defines $\pi_n(\theta)$ in a similar fashion.

Mis-specification

This framework facilitates inference in *mis-specified* models.

Suppose

$$\ell(\mathbf{y}, \theta) = -\log f(\mathbf{y}; \theta) - \log \pi_0(\theta)$$

represents a model that we do not think is the data generating model, but some other model in which we are interested.

- Why would we ever want to do this ?
- What does the “prior” term mean ?

Mis-specification

More generally, if $f(y; \theta)$ is acknowledged to be mis-specified

- what use is learning about θ ?;
- strict Bayesians may object to specifying a prior on θ .