Bayesian Theory with Applications

Jukka Corander
Department of Mathematics and statistics
University of Helsinki
Finland

Bruno de Finetti (1974,Theory of Probability):

”The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense *determined*, or known by other people, and so on, is of no consequence.”

Rissanen (1987, p.223), "As in Bayesian theory the class of models is not intended to include any "true" distribution for the data, but rather is only regarded as a language in which the properties of the data are to be expressed. This is a minimum requirement for any kind of learning, for how can we find regular features in the data unless we can describe them."

Probabilistic reasoning

These slides are primarily based on a compilation of material from the books Bernardo & Smith (1994), O'Hagan (1994), Schervish (1995), as well as additional material from lecturer's own research and other sources.

Handling uncertainty is undoubtedly a major part of all human activities, both scientific and non-scientific ones.

We have to make decisions and inference in situations where direct knowledge is not available to us.

Subjective probability, concerns the judgements of a given person, conveniently called You, about uncertain events or propositions.

The term *random quantity* is here used to signify a numerical entity whose value is uncertain.

The term *probability measure* or *distribution* (P) will be used in a rather loose manner (to avoid technicalities) to describe the way in which probability is "distributed" over the possible values of a random quantity.

When the probability distribution concentrates on a countable set of values, X is called a *discrete* random quantity, and we have the probability mass function $p(x) = P(X = x)$.

For *continuous* random quantities we have the regular *density* function representation $P(X \in B) = \int_B p(x)dx$.

Thus, to keep notation simple, $p(\cdot)$ is used both for mass and density functions.

Example 1 Thumbtack tossing. *Consider an old-fashioned thumbtack, which is of metal with a round curved head, rather than with a colored plastic one. The thumbtack will be tossed onto a soft surface (in order not damage it), while we keep track of whether it comes to stop with the point up or point down. In the absence of any information to distinguish the tosses or to suggest that tosses occurring close together in time are any more or less likely to be similar to or different from each other than those that are far apart in time, it seems reasonable to treat the different tosses symmetrically. We might also believe that although we might only toss the thumbtack a few times, if were to toss it many more times, the same judgement of symmetry would continue to apply to the future tosses.*

Under the above conditions, it is traditional to model the outcomes of the individual tosses as independent and identically distributed (IID) Bernoulli random quantities with $X_i = 1$ meaning that toss i is point up and $X_i = 0$ meaning that toss i is point down.

In the frequentist framework, one invents a **parameter**, say θ, which is assumed to be a fixed value in $[0, 1]$ not yet known to us (see the remark below).

Then one says that the X_i are IID with $P(X_i = 1) = \theta$. The so called **likelihood** function of a sequence of n tosses will under this assumption take the form

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1 - \theta)^{1-x_i} \tag{1}$$

which is the joint distribution of the observed values x_i conditional on θ.

The value of θ maximizing this function is the relative frequency of observing tosses point up, that is $\sum_{i=1}^{n} x_i/n$.

Given an observed sequence, our best guess of the probability of observing point up in the next toss, equals the relative frequency as well.

You immediately see what happens under scarce information, for instance, when the only two recorded tosses we have available are point down.

Given the earlier description of our simple thumbtack tossing problem, the assumptions made in the above frequentist approach (IID and fixed unknown θ) may appear unnecessary stringent.

In fact, this is remarkably true.

To derive a subjective probabilistic description of the behavior of the tosses, we need a minimal assumption of symmetry, called **exchangeability**.

Recall that we considered the information to be obtained from any one toss in exactly the same way we would consider the information from any other toss.

Similarly, we would treat the information to be obtained from any two tosses in exactly the same way we would consider the information from any other two tosses, regardless of where they appear in our sequence of tosses.

The same argument continues to apply to any subsequence of tosses.

These remarks of symmetry informally define the concept of exchangeability, which lies in the heart of the subjective probability description.

The concept and its generalization will be investigated formally later on.

As You might already have guessed, subjective probability description of the current situation, will require Your probabilistic description about the uncertainty related to the tosses (this will considered after the introduction of some formal concepts).

Remark 1 Meaning of the parameter in the thumbtack tossing problem. *A great deal of controversy in statistics arises out of the question of the meaning of such parameters as in the above example. De Finetti (1974) argues persuasively that one need not assume the existence of such things. Sometimes they are just assumed to be undefined properties of the experimental setup which magically make the outcomes behave according to our probability models. Sometimes they are defined in terms of the sequence of observations themselves (such as limits of relative frequencies). The last one is particularly troublesome because the sequence of observations does not yet exist and hence the limit of relative frequency cannot be a fixed value yet.*

From the above example we see the close connection between frequency probability and so called classical inference, because the latter requires the data to be repeatable.

An unbiased estimator, for instance, is defined to have expected value equal to the parameter being estimated.

Such statement is conditional on the parameter taking a fixed but unknown value, while the data are imagined as repeatable.

Typically, experimental data is thought to be repeatable, thus having frequency probability distribution, while parameters governing the data behavior in such framework are considered unique and unrepeatable.

Next example describes a situation where the frequency probability and classical inference seem to provide a distorted view of uncertainty.

Example 2 Evolution of species. *Consider a group of placental mammal species that are currently living on Earth. Evolutionary biologists working on the field of phylogenetics (see Felsenstein, 2004), wish to reconstruct the course of evolution among these species, by considering bits of DNA sequences sampled from individuals representing each of the species. A mathematical model in a form of stochastic process can be used to describe the evolution in terms of DNA site mutations. Unknown parameters in such a model typically involve a combinatorial object called tree, along with the branch lengths of the tree. The former describes the relationships between the species by stating an explicit neighborhood structure among them, while the latter represents the time that has evolved between speciation events. It is clear that only one evolution has taken place, it just happens to be unknown to us. Frequentist approach to uncertainty assessment about such trees would steer our thinking towards an idea of replicated earths where different evolutions have taken place - a kind of an odd perspective.*

Example 3 Meltdown accident of a nuclear reactor. *In engineering applications of reliability theory one often needs to consider the probability of an extremely rare event, such as a catasrophe. For instance, in US governmental regulations there is a statement that a power company aiming to produce electricity in a nuclear power plant, has to demonstrate that the probability of a critical meltdown accident at the plant will be less than one to a million at any time. It seems rather difficult to consider such an event from the frequentist point of view, through an imagined parameter that magically makes systems to break down every now and then, as we just keep track on them during a sufficiently long period of time. Note that two arbitrary observers of a power plant might have very different subjective probabilities about the plausibility of an accident. For instance, one of them might know that the responsible operating engineers happen to be drunk in the control room every night, and happily ignore their duties.*

The subjective view of probability is a much more powerful tool than the frequentist view, since it enables us to describe uncertainty in an arbitrary situation, by not restricting us to consider things imagined to be repeatable.

This, of course, is conditional on the fact that we have the mental skills of putting up the subjective probability description for the uncertainty we have at hand, which is sincerely difficult in many cases.

However, it should not be surprising that an approach which is fundamentally superior, requires more effort to be implemented.

In addition to the subjective probability, there exists also another prominent definition of probability in which a degree of belief in a proposition is considered.

This view is called the *logical* probability.

Recall that a subjective probability is a measure of *one* person's degree of belief.

Another person may have a different degree of belief in the same proposition, and so have a different probability.

The only constraint is that a single person's probabilities should not be inconsistent, and therefore they should obey all the axioms of probability.

We would expect two people's degrees of belief in a proposition to differ if they have different information, and in practice two people will never have exactly the same information.

The question does not arise, therefore, whether two people with identical information might have different subjective probabilities, or whether there is a unique degree of belief implied by a given body of information.

The latter view is taken in the theory of logical probability, which is then concerned with trying to identify logical degrees of belief.

Proponents of logical probabilities see them as extending the theory of logic.

In logic, a body of information may imply either the truth or falsehood of a given proposition, or may be insufficient to establish either truth or falsehood.

Logical probability is a measure of degree of implication when the information does not suffice to prove a proposition true or false.

Subjective and logical probabilities exist for any proposition, given any information.

Note that, a proposition may assert that a certain 'event' occurs, but the notion of a proposition is much more general than that of an event.

One difference between logical and subjective probabilities lies in propositions which are theoretically provable, such as the proposition that the four-hundredth digit in the decimal expansion of π is zero.

The logical probability of this proposition must be zero or one, and cannot be determined without computing the decimal expansion as far as the four-hundredth digit.

Most adherents of subjective probability would allow the individual to have a probability strictly between zero and one, if the decimal expansion of π is not immediately available. A common reaction would be to assign equal degrees of belief in the four-hundredth digit being 0, 1, 2, 3, 4, 5, 6, 7, 8 or 9, and hence to give the proposition of it being a zero a probability of one-tenth.

The Bayesian reasoning requires **prior** probabilities to be given explicit values.

A logical prior probability must be the unique value logically implied by the available prior information.

Unfortunately, for almost every kind of prior information this value cannot be found; the necessary theory simply does not exist.

The exception is the case where prior information is non-existent or, more accurately, where there is no prior information which is relevant to ”a thing like the parameter θ in the first example”.

For this case there exists a body of mathematical theory which can be used to construct prior distributions.

However, it is a matter of contention whether such a state of complete prior ignorance can exist.

Nevertheless, many proponents of subjective probability, even if they do not accept a true state of complete ignorance, will adopt these logical probabilities as approximations whenever prior information is very weak.

Bayesian reasoning in a nutshell

We examine first the Bayesian approach to probabilistic information processing in the simplest form. More comprehensive treatment will be given later.

Consider two events, A and B. From the identity

$$P(A)P(B|A) = P(A, B) = P(B)P(A|B) \tag{2}$$

we can form the simplest form of Bayes' theorem,

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \tag{3}$$

This formula can be interpreted as follows:

We are interested in the event B, and begin with an initial, *prior* probability $P(B)$ for its occurrence.

We then observe the occurrence of A.

The proper description of how likely B is when A is known to have occurred is the *posterior* probability $P(B|A)$.

Bayes' theorem can be understood as a formula for updating from prior to posterior probability, the updating consisting of multiplying by the ratio $P(A|B)/P(A)$.

It therefore describes how a probability changes as we learn new information.

Observing the occurrence of A will increase the probability of B if $P(A|B) > P(A)$. Using the law of total probability, we get

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) \tag{4}$$

where B^c denotes the complement of B ($P(B^c) = 1 - P(B)$). Further, we have

$$P(A|B) - P(A) = \{P(A|B) - P(A|B^c)\}P(B^c). \tag{5}$$

Assuming that $P(B^c) > 0$ (otherwise B is a certain event, and its probability would not be of interest), $P(A|B) > P(A)$ if and only if $P(A|B) > P(A|B^c)$.

Typically, the simple form of Bayes' theorem is given in a more general version.

Let $B_1, B_2, ..., B_n$ be a set of mutually exclusive and exhaustive events.

Then we have simple generalization of (3)

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i)P(A|B_i)}{P(A)} \qquad (6) \\ &= \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)} \end{aligned}$$

One can think of the different events B_i as a set hypotheses, one and only one of which is true (if hypothesis i is true we say that event B_i occurs).

Observing event A changes the prior probabilities B_i to posterior probabilities $P(B_i|A)$.

Notice that the posterior probabilities sum up to one (only one hypothesis is true).

The denominator $P(A)$ in (6) is a weighted average of the probabilities $P(A|B_i)$, the weights being the prior probabilities $P(B_i)$ (which sum to one).

The occurrence of A increases the probability of B_i if $P(A|B_i)$ is greater than the average of all $P(A|B_i)$'s.

The hypothesis whose probability is increased most by A (in the sense of being multiplied by the largest factor) is the one for which $P(A|B_i)$ is highest.

The probabilities $P(A|B_i)$ in (6) are known as *likelihoods*.

Specifically, $P(A|B_i)$ is the likelihood of B_i given by A.

The primitive notion, that hypotheses given greater likelihood by A should somehow have higher probability when A is observed to occur, has a compelling logic which is clearly understood from the following examples (due to Anthony O'Hagan).

Example 4 *I observe through my window a tall, branched thing with green blobs covering its branches: why do I think it is a tree? Because that is what trees generally look like. I do not think it is a man because men rarely look like that. Converting into formal notation,* A *is the event that I see a tall, branched thing partially covered in small green things,* B_1 *is the event that it is a tree,* B_2 *that it is a man, and* B_3 *that it is something else. The statement that ’trees generally look like that’ implies that* $P(A|B_1)$ *is close to one, whereas ’men rarely look like that’ means that* $P(A|B_2)$ *is close to zero. I convert these facts into a belief that the object is far more likely to be, i.e. has a much higher posterior probability of being, a tree than a man.*

Example 5 *A less extreme example occurs when I hear a piece of music but do not know its composer. I decide that the music is more probably Beethoven than Bach because, to me, it sounds more like the music Beethoven typically composed. That is, Beethoven has higher likelihood because a Beethoven composition is more likely to sound like this, and this suggests a higher probability for Beethoven.*

This primitive notion of likelihood underlies one of our most natural thought processes.

However, likelihood is not the only consideration in this reasoning.

Example 6 *Example 4 continued. There are other things that might look like a tree, particularly at a distance. I might be seeing a cardboard replica of a tree. This hypothesis would have essentially the same likelihood as the hypothesis that it is a tree, but it is not a hypothesis that I seriously entertain because it has a very much lower prior probability.*

Bayes' theorem (in the simple form in (6)) is in complete accord with the natural reasoning in Examples 4 and 6.

The posterior probabilities of the various hypotheses are in proportion to the products of their prior probabilities and their likelihoods.

Bayes' theorem thus combines two sources of information: the prior information is represented by the prior probabilities, the new information A is represented by the likelihoods, and the posterior probabilities represent the totality of this information.

Later on, when more probabilistic machinery has been introduced, we shall see at a general level that the Bayesian approach provides a very natural information processing system in empirical learning.

Example 7 Medical diagnosis. *In simple problems of medical diagnosis, Bayes' theorem often provides a particularly illuminating form of analysis of the various uncertainties involved. For simplicity, let us consider the situation where a patient may be characterized as belonging either to state* H_1*, or to state* H_2*, representing the presence or absence, respectively, of a specified disease. Let us further suppose that* $P(H_1)$ *represents the prevalence rate of the disease in the population to which the patient is assumed to belong, and that further information is available in the form of the result of a single clinical test, whose outcome is either positive (suggesting the presence of the disease and denoted by* $D = T$*), or negative (suggesting the absence of the disease and denoted by* $D = T^c$*).*

Example 8 Medical diagnosis continued. *The quantities* $P(T|H_1)$ *and* $P(T^c|H_2)$ *represent the* **true positive** *and* **true negative** *rates of the clinical test (often referred to as the test sensitivity and test specificity, respectively) and the systematic use of Bayes' theorem then enables us to understand the manner in which these characteristics of the test combine with the prevalence rate to produce varying degrees of diagnostic discriminatory power. In particular, for a given clinical test of known sensitivity and specificity, we can investigate the range of underlying prevalence rates for which the test has worthwhile diagnostic value.*

Example 9 Medical diagnosis continued. *As an illustration of this process, let us consider the assessment of the diagnostic value of stress thallium-201 scintigraphy, a technique involving analysis of Gamma camera image data as an indicator of coronary heart disease. On the basis of controlled experimental study, Murray et al. (1981) concluded that* $P(T|H_1) = 0.900$, $P(T^c|H_2) = 0.875$ *were reasonable orders of magnitude for the sensitivity and specificity of the test.*

Insight into the diagnostic value of the test can be obtained by plotting values of $P(H_1|T), P(H_1|T^c)$ *against* $P(H_1)$, *where*

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}, \tag{7}$$

for $D = T$ *or* $D = T^c$. *A graphical representation is given in Figure 1.1.*

Example 10 Medical diagnosis continued. *As a single, overall measure of discriminatory power of the test, one may consider the difference* $P(H_1|T) - P(H_1|T^c)$*. In cases where* $P(H_1)$ *has very low or very high values (e.g. for large population screening or following individual patient referral on the basis of suspected coronary disease, respectively), there is limited diagnostic value in the test. However, in clinical situations where there is considerable uncertainty about the presence of coronary heart disease, for example,* $0.25 \leq P(H_1) \leq 0.75$*, the test may be expected to provide valuable diagnostic information.*

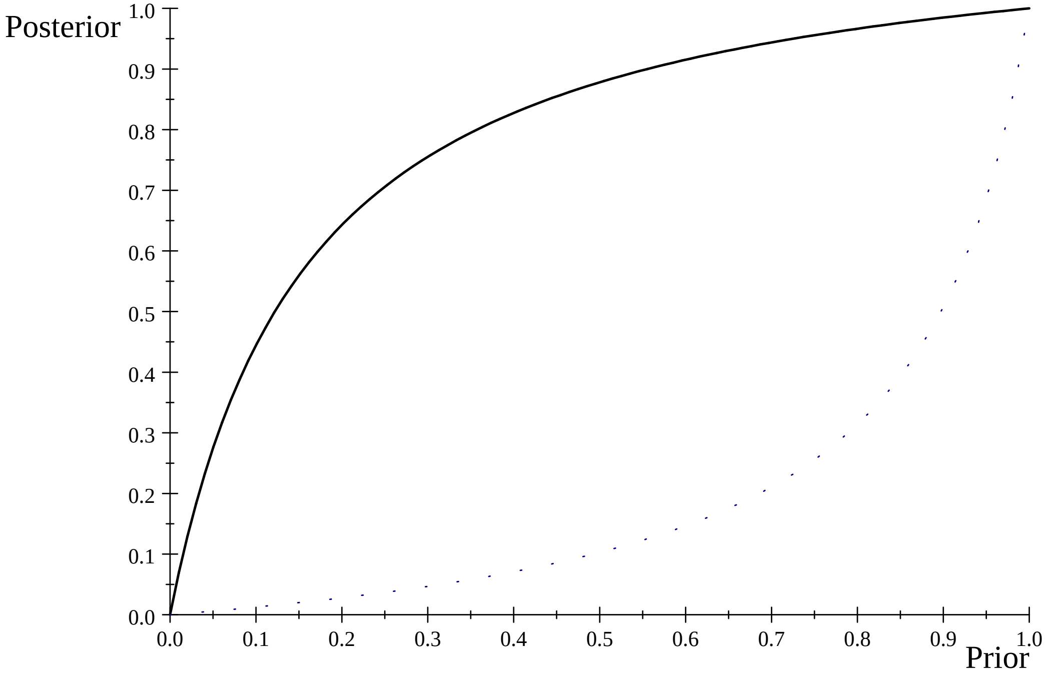


Figure 1.1. Values of $P(H_1|T)$ (thick solid line) and $P(H_1|T^c)$ (thin dotted line) against $P(H_1)$ (on x-axis).

One further point about the terms prior and posterior is worth emphasizing.

They are not necessarily to be interpreted in a chronological sense, with the assumption that "prior" beliefs are specified first and then later modified into "posterior" beliefs.

In any given situation, the particular order in which we specify degrees of belief and check their coherence is a pragmatic one.

Thus, some assessments seem straightforward and we feel comfortable in making them directly, while we are less sure about other assessments and need to approach them indirectly via the relationships implied by coherence.

It is true that the natural order of assessment does coincide with the "chronological" order in a number of practical applications, but is important to realize that this is a pragmatic issue and not a requirement of the theory.

As seen above, Bayes' theorem characterizes the way in which beliefs about "hypotheses" $H_i, i = 1, ..., n$, (earlier denoted by events B_i) are revised in the light of new observations D.

In many cases we receive data in successive stages, so that the process of revising beliefs is sequential.

As a simple illustration of this process, let us suppose that data are obtained on two stages, which can be described by real-world events D_1 and D_2.

Now, revision of the beliefs on the basis of the first piece of data D_1 is described by

$$P(H_i|D_1) = P(D_1|H_i)P(H_i)/P(D_1), i = 1, ..., n \tag{8}$$

When it comes to the further, subsequent revision of beliefs in the light of D_2, the likelihoods and prior probabilities to be used in the Bayes' theorem are now $P(D_2|H_i \cap D_1)$ and $P(H_i|D_1)$, for $i = 1, ..., n$, respectively, since all judgements are now conditional on D_1.

We thus have

$$P(H_i|D_1 \cap D_2) = \frac{P(D_2|H_i \cap D_1)P(H_i|D_1)}{P(D_2|D_1)} \tag{9}$$

where $P(D_2|D_1) = \sum_{i=1}^{n} P(D_2|H_i \cap D_1)P(H_i|D_1)$.

From an intuitive standpoint, we would obviously anticipate that coherent revision of initial belief in the light of combined data $D_1 \cap D_2$ should not depend on whether D_1 and D_2 were analyzed successively or in combination.

This is easily verified by substituting the expression for $P(H_i|D_1)$ into the expression for $P(H_i|D_1 \cap D_2)$, by which we get

$$\frac{P(D_2|H_i \cap D_1)P(D_1|H_i)P(H_i)}{P(D_2|D_1)P(D_1)} = \frac{P(D_1 \cap D_2|H_i)P(H_i)}{P(D_1 \cap D_2)} \tag{10}$$

the latter being the direct expression for $P(H_i|D_1 \cap D_2)$ when $D_1 \cap D_2$ is treated as a single piece of data.

This procedure generalizes to any number of sequential stages where data are observed.

Independence concepts and graphs

Since the concept of independence (conditional and marginal) is very fundamental in statistical modelling, it will be useful to study some of its properties.

Also, the connection between mathematical objects called **graphs** and probabilistic independence is very useful in Bayesian modelling.

For instance, the popular WinBugs software for Bayesian modelling exploits this connection to a large extent.

Let X and Y be random quantities and P their joint distribution.

The relation $X \perp Y$ denotes that, under P, X and Y are (pairwise) independent.

For measurable sets A, B, this relation implies that

$$\begin{aligned} X &\perp Y \\ &\Leftrightarrow \\ P(X &\in A, Y \in B) = P(X \in A)P(Y \in B) \\ &\Leftrightarrow \\ P(X &\in A|Y \in B) = P(X \in A) \end{aligned}$$

Consider now the random quantities X, Y, Z their joint distribution P.

The relation $X \perp Y|Z$ denotes that, under P, X and Y are conditionally independent given any realization of Z.

For measurable sets A, B, this relation implies that

$$\begin{aligned} X &\perp Y|Z \\ &\Leftrightarrow \\ P(X &\in A, Y \in B|Z) = P(X \in A|Z)P(Y \in B|Z) \\ &\Leftrightarrow \\ P(X &\in A|Y, Z) = P(X \in A|Z) \end{aligned}$$

An important result which strengthens the interpretation of conditional independences is the following.

Let the discrete random quantities X, Y, Z have a strictly positive probabilities for all outcomes according to their joint distribution P.

Then, the conditional independence relation will satisfy the following marginal independence condition

$$X \perp Y|Z, X \perp Z|Y \Rightarrow X \perp (Y, Z)$$

There exists a general recursive representation of joint distributions, which is often useful.

Assume there are k random quantities in a set V.

These are labelled in some order with integers as $1, ..., k$.

Let the set $V(i) = \{1, ...i\}$ denote the set of i and its predecessors, for any $i = 1, ..., k$ (with respect to the fixed order).

The joint distribution can be specified in terms of the *recursive factorization*

$$P_V = P_1 P_{2|1} \cdots P_{k-1|V(k-2)} P_{k|V(k-1)} \tag{11}$$

A *graph* is a mathematical object, formed by a pair $G = (V, E)$,

where V is finite set of *vertices* (nodes), and

E the *edge* set, which is a subset of the cartesian product $V \times V$.

Here vertices represent random quantities.

The graphs considered here contain no loops, so that $(v, v) \notin E$, for all $v \in V$.

In visualization of a graph, it is often convenient to use as labels of the elements of V the set of integers $\{1, ..., k\}$.

When both (v, v') and (v', v) belong to E for some distinct v and v', the edges are called *undirected*.

On contrary, when only one of these ordered pairs is in E, the edge is called *directed* (arc or arrow).

Vertices $\{v', v\}$ are *adjacent* if there is an edge connecting them, otherwise they are *non-adjacent*.

If a graph has only undirected edges it is called undirected, in which case it is more convenient to represent the edges as unordered pairs $\{v, v'\}$.

When all edges are directed the graph is called directed.

A *path* is a sequence of vertices for which there is an edge $(v, v') \in E$, for every pair of successive elements $\{v, v'\} \subset V$.

A path is also a *cycle* if the first and last elements of the sequence are equal.

In an undirected graph a cycle is *chordless* if *only* successive pairs of elements in it are adjacent.

A directed graph is *acyclic* if it does not contain any cycles, such graphs are typically called DAGs.

For directed edges it is relevant to consider the following two order-related concepts.

If $(v, v') \in E$, v is called a *parent* of v', and v' a *child* of v.

The set of parents of v is denoted by $pa(v)$.

The *ancestors* $an(v)$ of a vertex $v \in V$ are the vertices from which there is a path *leading to* v.

We now turn to relating formally the concept of conditional independence to graphs.

When a joint distribution P of V satisfies the structure specified by a DAG, we can strengthten the recursive factorization as

$$
\begin{aligned}
P_V &= P_1 P_{2|1} \cdots P_{k-1|V(k-2)} P_{k|V(k-1)} && (12)\\
&= \prod_{v \in V} P_{v|pa(v)} && (13)
\end{aligned}
$$

This means that v is conditionally independent of the remaining ancestors, given its parents.

This powerful property of DAGs enables the construction of joint distributions of very large node sets.

The DAGs are also a convenient (and currently fairly standard) way of communicating complex hierarchical statistical models.

Undirected graphs are also useful in the context of representing independence.

To be able to explore this, we need some further notations.

The *boundary* $bd(A)$ of $A \subseteq V$, is the set of vertices in $V \backslash A$ that have an undirected edge to vertices in A.

For a subset $A \subseteq V$, the subgraph induced by A is $G_A = (A, E_A)$ with $E_A = E \cap (A \times A)$.

A graph G is *complete* when all pairs of vertices are adjacent.

A *clique* is a subset $A \subseteq V$, for which G_A is complete and for any non-empty $B \subseteq V \backslash A$, $G_{A \cap B}$ is *not* complete (G_A is *maximally* complete).

A subset $C \subset V$ *separates* the disjoint subsets A and B, if for every pair $v \in A, v' \in B$, all paths from v to v' go through C.

For an undirected graph $G = (V, E)$ associated with a joint probability distribution P for V, there are various **Markov** properties P may satisfy.

(1) **Pairwise Markov property** : $v \perp v' | V \backslash \{v, v'\} \Leftrightarrow \{v, v'\} \notin E$

(2) **Local Markov property** : for all $v \in V$, $v \perp V \backslash (\{v\} \cup bd(v)) | bd(v)$

(3) **Global Markov property** : for any triple (A, B, C) of disjoint subsets of V such that C separates A from B in G, $A \perp B | C$

The strength of contraints imposed by different Markov properties varies in general, and is given by the following theorem.

Theorem 1 *For* **any** *undirected* G *and* P, *the markov properties satisfy* (3) $\Rightarrow$ (2) $\Rightarrow$ (1)

A fundamental part of the theory of undirected graphs is related to the decomposition of graphs into subgraphs.

The following definitions are essential in the development of further concepts.

Definition 1 *A partition $V = A \cup B \cup C$ of the vertex set of an undirected marked graph G forms a decomposition of G if*

(i) C separates A from B

(ii) G_C is complete

Definition 2 *An undirected marked graph G is decomposable if*

(i) G is complete or

(ii) There exists a decomposition (A, B, C) into decomposable subgraphs $G_{A\cup C}$ and $G_{B\cup C}$

A decomposable graph can be recursively split into its cliques by decompositions.

Let $\mathcal{C}(G)$ denote the class of cliques of a decomposable graph.

The class of separators $\mathcal{S}(G)$ (intersections of successive cliques) is obtained by a series of decompositions of G that leads into $\mathcal{C}(G)$.

A *triangulated* graph is such an undirected graph that it contains no chordless cycles of length four or larger.

An important graph theoretical result is stated below for triangulated graphs.

Theorem 2 *For an undirected graph G, the following conditions are equivalent*

(i) G is decomposable

(ii) G is triangulated

(iii) every minimal separator of a pair of vertices induces a complete subgraph

A particularly useful form of factorization is available for decomposable graphs, for which the joint distribution can be represented as

$$P = \frac{\prod_{c \in \mathcal{C}(G)} P_c}{\prod_{s \in \mathcal{S}(G)} P_s} \tag{14}$$

where P_a is the marginal distribution of $a \subset V$.

The is a close relation between undirected graphs and DAGs.

To investigate that we need som further concepts.

For a DAG G we define the *moral graph* G^m as the undirected graph obtained from G by:

(*i*) replacing all directed edges by undirected ones

(*ii*) for any node v in G, with non-adjacent parents $z \in pa(v)$ and $z' \in pa(v')$, inserting the edge $\{z, z'\}$ into G^m.

The term moral graph refers to the "marrying of parents".

The following theorem states a useful results for DAGs.

Theorem 3 *If a probability distribution satisfies a recursive factorization according to the DAG G, it factorizes with respect to the moral graph G^m and therefore shares the global Markov property of G^m.*

It should be noticed that the moral graph G^m on the whole vertex set may obscure certain marginal independences present in a DAG G. These can be deduced via the moral graphs on ordered segments of V, instead of the whole vertex set.

Usefulness of the graph concepts combined with Markov properties is illustrated with separate case studies.

Subjective probability modeling

In our framework probabilities are always *personal degrees of belief*, in that they are a numerical representation of an analyst's or decision maker's personal uncertainty relation between events.

Moreover, probabilities are always conditional on the information available.

It makes thus no sense to qualify the word probability with adjectives such as "objective", "correct" or "unconditional".

As clearly stated in de Finetti (1974), to be able to use probability calculus as a normative tool for the description of the characteristics of interest for random quantities, one *has to* express individual degrees of belief (*i.e.* subjective opinions), expressed as probabilities about the uncertainty involved in the considered situation.

That is, phrases such as "I don't know", "I can't" or "I don't want to" cannot be accepted as answers to the question concerning what one's beliefs are.

The failure to express these probabilities will lead us outside the Bayesian paradigm (in the stringent sense).

However, in the literature Bayesian paradigm is often understood more widely, including even cases where the subjective probabilities are replaced by formally derived functions (this aspect will be considered more in depth later).

Using generic notation we assume that the subjective degrees of belief correspond to the specification of the joint distribution $P(x_1, ..., x_n)$ of a set of random quantities $\mathbf{x} = x_1, ..., x_n$, represented by the joint density (or mass) function $p(x_1, ..., x_n)$.

This specification automatically leads, for $1 \leq m < n$, to the marginal joint density

$$p(x_1, ..., x_m) = \int p(x_1, ..., x_n) dx_{m+1} \dots dx_n \tag{15}$$

and the joint density of $\mathbf{y} = x_{m+1}, ..., x_n$ (thought as yet unobserved), conditional on having observed the particular values of $\mathbf{z} = x_1, ..., x_m$, is

$$p(x_{m+1}, ..., x_n | x_1, ..., x_m) = \frac{p(x_1, ..., x_n)}{p(x_1, ..., x_m)} \tag{16}$$

A *predictive probability model* for random quantities can be defined according to the following.

Definition 3 Predictive probability model. *A predictive model for a sequence of random quantities $x_1, x_2, ...$ is a probability measure P, which specifies the joint belief distribution for any subset of $x_1, x_2, ...$.*

Consider now a sequence $x_1, x_2, ...$ under the assumption of a predictive model stating that for any n the joint density is given by

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i) \tag{17}$$

This model thereby states that the uncertain quantities are independent.

If we now consider the conditional density for $1 \leq m < n$, it takes the form

$$p(x_{m+1}, ..., x_n | x_1, ..., x_m) = p(x_{m+1}, ..., x_n) \tag{18}$$

meaning that we cannot learn from experience within this sequence of interest.

In other words, past data provide us with no additional information about the possible outcomes of future observations in the sequence.

A predictive model specifying such independence is clearly inappropriate in contexts where we believe that the successive accumulation of data will provide increasing information about future events.

Thus, in most cases a useful predictive model, *i.e.* the structure of $p(x_1, ..., x_m)$, ought to encapsulate some form of dependence among the individual random quantities.

In general, there are a vast number of possible subjective assumptions about the form of such dependencies, and here we are able to consider some commonly used canonical forms.

Suppose that, in thinking about $P(x_1, ..., x_n)$, the joint degree of belief distribution for a sequence of random quantities $x_1, ..., x_m$, an individual makes the judgement that the subscripts or the labels identifying the individual random quantities, are "uninformative".

The uninformativeness is in the sense that the same marginal distribution would be specified for all possible singletons, pairs, triples etc., regardless of which labels were happened to be picked from the original sequence (recall the thumbtack tossing in Example 1).

This leads us to the concept of exchangeability, formally defined below.

Definition 4 Exchangeability. *Random quantities $x_1, ..., x_n$ are said to be (finitely) exchangeable under a probability measure P when the corresponding joint belief distribution satisfies*

$$p(x_1, ..., x_n) = p(x_{\pi(1)}, ..., x_{\pi(n)})$$

for an arbitrary permutation π of the labels $\{1, ..., n\}$. Further, an infinite sequence $x_1, x_2, ...$ is said to be infinitely exchangeable when every finite subsequence is finitely exchangeable.

For example, suppose that $x_1, ..., x_{100}$ are exchangeable.

It follows from the above definition that they all have the same marginal distribution.

Also, (x_1, x_2) has the same joint distribution as (x_{99}, x_1), and (x_5, x_2, x_{48}) has the same joint distribution as (x_{31}, x_{32}, x_{33}), and so on.

The notion of exchangeability involves a judgement of complete symmetry among all the observables $x_1, ..., x_n$ under consideration.

Clearly, in many situations this might be too restrictive an assumption, even though a partial judgement of symmetry is present, which should be evident from the following example.

Example 11 Tossing with different thumbtacks. *Consider a scenario which is similar to that of Example 1, except that we make* n_i, $i = 1, ..., k$, *tosses with thumbtacks of different material. For instance, the first thumbtack is made of metal, the second of plastic, the third of kevlar and so on. We might be involuntary to describe a sequence of observations under this scenario using the complete symmetry assumption leading to exchangeability. On the other hand, as before it should be reasonable to treat tosses made with the same thumbtack as exchangeable.*

We now formally treat the subjective modeling problem of an infinitely exchangeable sequence of $0 - 1$ (binary) random quantities (say, thumbtack tosses) $x_1, x_2, ...$ with $x_i = 0$ or $x_i = 1$, for all $i = 1, 2, ...$.

Theorem 4 Representation theorem for binary random quantities. *If $x_1, x_2, ...$ is an infinitely exchangeable sequence of binary random quantities with probability measure P, there exists a distribution function Q such that the joint mass function $p(x_1, ..., x_n)$ for $x_1, ..., x_n$ can be written as*

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta) \tag{19}$$

where

$$Q(\theta) = \lim_{n\to\infty} P[y_n/n \le \theta]$$

and $y_n = \sum_{i=1}^n x_i$, $\theta = \lim_{n\to\infty} y_n/n$.

The interpretation of this representation theorem is of profound significance from the point of view of subjectivist modeling philosophy. It is *as if*:

- The x_i are judged to be independent, Bernoulli random quantities conditional on a random quantity θ.
- θ is itself assigned a probability distribution $Q(\theta)$.
- By the strong law of large numbers $\theta = \lim_{n\to\infty} y_n/n$, so that Q may be interpreted as "beliefs about the limiting frequency of 1's".

What the above says is that, under the assumption of exchangeability, we may act *as if*, conditional on θ, the quantities $x_1, ..., x_n$ are a *random sample* from a Bernoulli distribution with parameter θ which corresponds to the joint sampling distribution (the *likelihood*)

$$p(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{20}$$

where the parameter θ is given a *prior distribution* $Q(\theta)$.

Notice that under this interpretation the prior states beliefs about what we would anticipate observing as the limiting relative frequency.

Further, the assumption of exchangeability in the current framework considerably limits via Theorem 4 our alternatives in the specification of a predictive probability model.

Any choice must be of the form given by (19), where we have the freedom of choosing the subjective beliefs about θ.

By ranging over all possible choices of the prior $Q(\theta)$, we build all possible predictive probability models for the current framework.

We have thus established a justification for the conventional model building procedure of combining a likelihood and a prior.

The likelihood is defined in terms of an assumption of conditional independence of the observations given a parameter.

This, and its associated prior distribution, acquire an operational interpretation in terms of a limiting average of observables (here limiting frequency).

In many applications involving binary random quantities, we may be more interested in a summary random quantity, such as $y_n = x_1 + \cdots + x_n$, than in the individual sequences of x_i's.

The representation $p(y_n)$ follows easily from (19), since

$$p(y_n) = \binom{n}{y_n} p(x_1, ..., x_n),$$

for all $x_1, ..., x_n$ such that $x_1 + \cdots + x_n = y_n$.

We thus get

$$p(y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1 - \theta)^{n - y_n} dQ(\theta)$$

This provides a justification, when expressing beliefs about y_n, for acting *as if* we have a binomial likelihood with a prior distribution $Q(\theta)$ for the binomial parameter θ.

The Bayesian learning process in this simple situation is compactly represented by the following corollary.

Corollary 5 Corollary to the Representation theorem for binary random quantities. *If $x_1, x_2, ...$ is an infinitely exchangeable sequence of binary random quantities with probability measure P, the conditional probability function $p(x_{m+1}, ..., x_n | x_1, ..., x_m)$ for $x_{m+1}, ..., x_n$ given $x_1, ..., x_m$, has the form*

$$\int_0^1 \prod_{i=m+1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta | x_1, ..., x_m) \tag{21}$$

where

$$dQ(\theta | x_1, ..., x_m) = \frac{\prod_{i=1}^{m} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta)}{\int_0^1 \prod_{i=1}^{m} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta)}$$

and

$$Q(\theta) = \lim_{n\to\infty} P[y_n/n \leq \theta]$$

and $y_n = \sum_{i=1}^{n} x_i$, $\theta = \lim_{n\to\infty} y_n/n$.

We thus see that the basic form of representation of beliefs does not change.

All that has happened, expressed in conventional terminology, is that the *prior* distribution $Q(\theta)$ for θ has been revised into the *posterior* distribution $dQ(\theta|x_1, ..., x_m)$.

The conditional probability function $p(x_{m+1}, ..., x_n|x_1, ..., x_m)$ is called the *posterior predictive* probability function.

This provides the basis for deriving the conditional predictive distribution of a generic random quantity defined in terms of the future observations.

In a more general setup the representation theorem states for an infinitely exchangeable sequence of real valued quantities $x_1, x_2, ...$ with probability measure P, that there exists a probability measure Q over the space $\mathcal{Q}$ of all distribution functions for the observable quantity, such that the joint distribution function of $x_1, ..., x_n$ can be written as

$$P(x_1, ..., x_n) = \int_{\mathcal{Q}} \prod_{i=1}^{n} F(x_i) dQ(F) \tag{22}$$

where

$$Q(F) = \lim_{n \to \infty} P(F_n) \tag{23}$$

where F_n is the empirical distribution function defined by $x_1, ..., x_n$.

Thus, we may act *as if* we have independent observations $x_1, ..., x_n$ conditional on F, which is an unknown distribution function playing the role of an infinite-dimensional parameter.

The belief distribution Q has in this case the interpretation of what we believe the empirical distribution function F_n would look like for a "large" number of observations.

This result can be analogously extended to a finite dimensional Euclidean space for vector valued random quantities.

If, in particular, our beliefs are such that the distribution function F can be defined in terms of a finite-dimensional parameter $\boldsymbol{\theta}$, the joint density of our observations can be written as

$$p(x_1, ..., x_n) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) \tag{24}$$

where $p(\cdot|\boldsymbol{\theta})$ is the density function corresponding to the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By taking a step yet further, and letting $p(\boldsymbol{\theta})$ correspond to the density representation of $Q(\boldsymbol{\theta})$, we obtain

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{25}$$

From the above we may deduce that

$$
\begin{aligned}
p(x_{m+1},...,x_n|x_1,...,x_m) &= \frac{\int \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \qquad (26)\\
&= \frac{\int \prod_{i=m+1}^{n} p(x_i|\boldsymbol{\theta}) \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}
\end{aligned}
$$

where

$$
\frac{\prod_{i=1}^{m} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = p(\boldsymbol{\theta}|x_1,...,x_m) \qquad (27)
$$

so that

$$
p(x_{m+1},...,x_n|x_1,...,x_m) = \int \prod_{i=m+1}^{n} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1,...,x_m)d\boldsymbol{\theta} \qquad (28)
$$

The relation in (27) is just *Bayes' theorem*, which expresses the posterior density for $\boldsymbol{\theta}$ in the context of parametric model for $x_1,...,x_m$ given $\boldsymbol{\theta}$.

By using the more compact notations about the "future" $\mathbf{y}$ and the "current" observations $\mathbf{z}$, we see that

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (29) \\ p(\mathbf{y}|\mathbf{z}) &= \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta} \\ p(\boldsymbol{\theta}|\mathbf{z}) &= \frac{p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{z})} \end{aligned}$$

In particular the role of Bayes' theorem is identified as a coherent learning step about the unobservables when we pass from $p(\mathbf{z})$ to $p(\mathbf{y}|\mathbf{z})$.

Illustrations with probabilistic classifiers

To put the above theoretical franework into an applied context we consider probabilistic classification.

Assume we wish to use Bayes' theorem and exchangeability to filter out spam messages from email inflow.

Let's first have a look at the solution derived by maximum likelihood estimation combined with Bayes' theorem.

Let the vector $y = (y_1, ..., y_d)$ represent the information we have extracted from a new email message represented by a sequence $\mathbf{s} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T\}$ of strings (words).

Each y_j is an indicator variable for a specific word (say 'HOT') being present in the message, i.e. $y_j = 1$ if 'HOT'$\in \mathbf{s}$, and $y_j = 0$ otherwise.

Assume we have in total defined d different words in a similar fashion.

Assume that we have previously examined a set of messages and determined whether they represent spam (n_1 messages) or not (n_2 messages).

These two classes of n_1 and n_2 messages will be indexed by $c = 1$ and $c = 2$, respectively.

We can call the previously examined messages *training data*.

Information in each of these messages is condensed in the binary vector $x_i = (x_{i1}, ..., x_{id})$, where the element x_{ij} is defined equivalently to the above definition for y_j, i.e. indicates the presence/absence of a specific word within the message.

Define p_{cj} as the probability of observing the word j in an arbitrary message sampled from the population represented by class c.

In our previous notation we can define $p_{cj} = \theta_{cj}$.

Assume all d words appear in a message *independently of each other, conditional on* $\theta_{cj}, j = 1, ..., d$.

Using maximum likelihood approach we may now estimate θ_{cj} separately for each j using the frequency data from the n_1, n_2 training messages.

This gives

$$\hat{\theta}_{1j} = n_1^{-1} \sum_{i=1}^{n_1} x_{ij}, \tag{30}$$

and

$$\hat{\theta}_{2j} = n_2^{-1} \sum_{i=1}^{n_2} x_{ij}, j = 1, ..., d. \tag{31}$$

The maximum likelihood approach has thus delivered us a characterization of the probability (likelihood) of the presence/absence pattern of words in any future message sampled from the corresponding class.

Let $z \in \{1, 2\}$ denote the unobserved event indicating from which class the message with information y was sampled.

The likelihood of $y = (y_1, ..., y_d)$ in class $c = 1$ now equals:

$$\begin{aligned} \hat{p}(y|z &= 1) = \hat{\theta}_{11}^{y_1}(1-\hat{\theta}_{11})^{1-y_1}\hat{\theta}_{12}^{y_2}(1-\hat{\theta}_{12})^{1-y_2}\cdots\hat{\theta}_{1d}^{y_d}(1-\hat{\theta}_{1d})^{1-y_d} \\ &= \prod_{j=1}^{d}\hat{\theta}_{1j}^{y_j}(1-\hat{\theta}_{1j})^{1-y_j}. \end{aligned} \tag{32}$$

Similarly, we get for the other class

$$\hat{p}(y|z = 2) = \prod_{j=1}^{d}\hat{\theta}_{2j}^{y_j}(1-\hat{\theta}_{2j})^{1-y_j}. \tag{33}$$

Let now $p(z = 1) = 1 - p(z = 2)$ be the *prior* probability that a message will come from class $c = 1$.

Corresponding probability is thus defined for $c = 2$.

Depending on the situation, we might wish to set $p(z = 1) = p(z = 2)$, or we might wish to use as $p(z = 1)$ the fraction of spam messages we anticipate to receive among the total inflow.

By feeding the above ingredients into Bayes' theorem we obtain the *posterior probability of the new message being spam:*

$$\hat{p}(z=1|y) = \frac{\hat{p}(y|z=1)p(z=1)}{\hat{p}(y|z=1)p(z=1) + \hat{p}(y|z=2)p(z=2)}. \tag{34}$$

What we just have defined is generally known as the *naive Bayes classifier* based on maximum likelihood rule.

We may wish use an additional ingredient based on decision theory, and flag messages as spam *only if* the posterior probability $\hat{p}(z=1|y)$ exceeds a pre-specified threshold, such as 0.9, instead of just using the rule which assigns messages to the class having highest posterior probability.

REMARK! Even if the word indicators are highly dependent, the above classifier may work surprisingly well.

Why could that be? Think about the situation where marginal probabilities $\hat{\theta}_{1j}, \hat{\theta}_{2j}$ are close to either 0 or 1. What happens to the approximation of the joint distribution of the elements in y?

The classifier we derived above does not follow strictly the principles of predictive learning as derived from axioms of probability, and therefore, it is not surprising that one may encounter problems under various circumstances.

Consider the situation where word j does not occur in class $c = 1$ and word k does not occur in class $c = 2$.

Then,

$$\begin{aligned} \hat{\theta}_{1j} &= 0 \qquad (35) \\ \hat{\theta}_{2k} &= 0. \end{aligned}$$

Consequently, for any message that contains both words j and k, the posterior probability equals zero for both classes, because the data are impossible under both $z \in \{1, 2\}$.

Such a situation is easily encountered if the amount of training data (n_1, n_2)

is small.

Instead of using the maximum likelihood approach presented above, we may derive a classifier starting from the principles of predictive modeling.

Warning! Take a deep breath before reading the section below ;)

Assume that the observed word patterns x_i in the messages are *unrestrictedly infinitely exchangeable* within class $c = 1$ (see Definitions 4.2, 4.3, and 4.13, and Propositions 4.2 and 4.18 in Bernardo & Smith, 1994).

This assumption implies that, if we combine any permutation of the message values $x_{1j}, ..., x_{n_c j}$, for a fixed $j = 1, ..., d$, with arbitrary corresponding permutations over the remaining word indicators, the same predictive probability mass function for the data $x_i, i = 1, ..., n_c$, is obtained. Furthermore, we also obtain a characterization of the sequences $x_{1j}, ..., x_{n_c j}$ in terms of the sufficient statistics, which equal the observed number of word j among the training messages in the class c.

Let $\boldsymbol{\theta}_c = (\theta)_{cj}, j = 1, ..., d$, be the vector of word frequencies in class c.

By assuming the unrestricted infinite exchangeability to hold in both classes we get an explicit expression (see below for details) for the predictive distribution for future class data:

$$\begin{aligned} p(y|x_1,...,x_{n_c}) &= \int_{\Theta_c} p(y|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c|x_1,...,x_{n_1})d\boldsymbol{\theta}_c \quad (36) \\ &= \int_{\Theta_c} p(y|\boldsymbol{\theta}_c)\frac{p(x_1,...,x_{n_c}|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c)}{\int_{\Theta_c} p(x_1,...,x_{n_c}|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c)d\boldsymbol{\theta}_c}, \quad (37) \end{aligned}$$

where $p(\boldsymbol{\theta}_c)$ is a joint prior distribution for the word frequencies in class c and $p(\boldsymbol{\theta}_c|x_1,...,x_{n_c})$ is the corresponding posterior obtained by conditioning on the observed training data.

Notice that the parameters in $\boldsymbol{\theta}_c$ are integrated out because they are of no interest in itself for this prediction problem, and the result is thus a simple consequence of applying probability axioms. It is fairly common in probabilistic prediction tasks that even if the parameters of a finite-dimensional model have an operational meaning like here, they themselves are not target of inference.

The contrast between (36) and the earlier predictive distribution based on maximum likelihood (32) becomes clear once considered from a probabilistic perspective.

The former addresses the predictive uncertainty by replacing the unknown parameters by point estimates, whereas the latter calculates predictive average likelihood with respect to the posterior distribution obtained by combining initial uncertainty (prior $p(\boldsymbol{\theta}_c)$) with the empirical information.

Thus, we see that from the probabilistic perspective the predictive distribution of y is necessarily an (infinite) mixture, where the variance is in general larger than in the ordinary model with fixed parameter values.

The limiting behavior of the predictive distribution as a function of the amount of training data is also illuminating and we will examine that explicitly after developing the expression (36) under specific prior assumptions.

Assume that for any component θ_{cj} in $\boldsymbol{\theta}_c$ the prior is specified as the Beta(α, β) distribution, which has the density

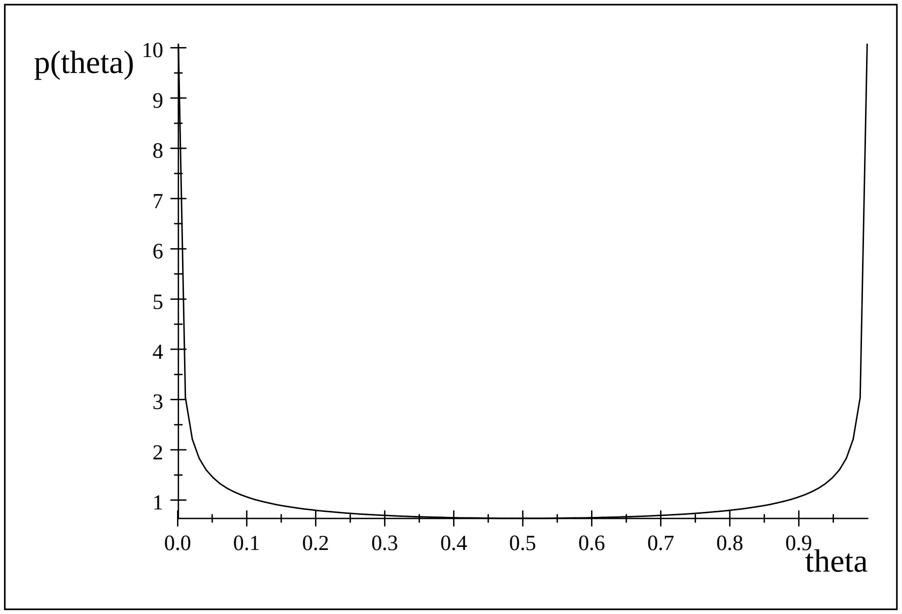
$$p(\theta_{cj}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_{cj}^{\alpha-1}(1-\theta_{cj})^{\beta-1}, \tag{38}$$

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

The assumption that α, β are not indexed by j reflects that our prior characterization of the uncertainty is exchangeable for all word frequencies. A wide range of different prior beliefs can be represented by choosing the hyperparameters α, β suitably.

The Beta family of distributions is conjugate prior for the Bernoulli and Binomial sampling models, which implies that the posterior belongs to the same family of distributions.

For instance, the choice $\alpha = \beta = 1/2$ reflects the belief that low and high word frequencies are more likely than intermediate ones, while still leading to a symmetric prior distribution. There lies also a theoretical motivation behind this particular prior and it can be derived using so called Jeffreys' and Perks' principles, which will be discussed later. The density of this distribution has the U-shape shown below.



The generalized exchangeability assumption implies that we consider the d word frequencies as conditionally independent and the corresponding joint prior distribution then becomes

$$\begin{aligned} p(\boldsymbol{\theta}_c) &= p(\theta_{c1}) \cdots p(\theta_{cd}) \qquad (39) \\ &= \prod_{j=1}^{d} p(\theta_{cj}) \\ &= \prod_{j=1}^{d} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_{cj}^{\alpha-1}(1-\theta_{cj})^{\beta-1}. \end{aligned}$$

Under the above prior the posterior distribution is analytically tractable and equals

$$
\begin{aligned}
p(\boldsymbol{\theta}_c|x_1,...,x_{n_c}) &= \frac{\prod_{j=1}^{d} p(x_{1j},...,x_{n_cj}|\theta_{cj})p(\theta_{cj})}{\prod_{j=1}^{d} \int_0^1 p(x_{1j},...,x_{n_cj}|\theta_{cj})p(\theta_{cj})d\theta_{cj}} \\
&= \frac{\prod_{j=1}^{d} \prod_{i=1}^{n_c} \theta_{cj}^{x_{ij}}(1-\theta_{cj})^{1-x_{ij}} p(\theta_{cj})}{\prod_{j=1}^{d} \int_0^1 \prod_{i=1}^{n_c} \theta_{cj}^{x_{ij}}(1-\theta_{cj})^{1-x_{ij}} p(\theta_{cj})d\theta_{cj}}
\end{aligned}
$$

$$= \frac{\prod_{j=1}^{d} \theta_{cj}^{\sum_{i=1}^{n_c} x_{ij}} (1-\theta_{cj})^{n_c - \sum_{i=1}^{n_c} x_{ij}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_{cj}^{\alpha-1} (1-\theta_{cj})^{\beta-1}}{\prod_{j=1}^{d} \int_0^1 \theta_{cj}^{\sum_{i=1}^{n_c} x_{ij}} (1-\theta_{cj})^{n_c - \sum_{i=1}^{n_c} x_{ij}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_{cj}^{\alpha-1} (1-\theta_{cj})^{\beta-1} d\theta_{cj}}$$

$$= \frac{\prod_{j=1}^{d} \theta_{cj}^{\alpha+\sum_{i=1}^{n_c} x_{ij}-1} (1-\theta_{cj})^{\beta+n_c - \sum_{i=1}^{n_c} x_{ij}-1}}{\prod_{j=1}^{d} \int_0^1 \theta_{cj}^{\alpha+\sum_{i=1}^{n_c} x_{ij}-1} (1-\theta_{cj})^{\beta+n_c - \sum_{i=1}^{n_c} x_{ij}-1} d\theta_{cj}}.$$

The above distribution is recognized as a product of d Beta distributions, where the jth element equals Beta$(\alpha + \sum_{i=1}^{n_c} x_{ij}, \beta + n_c - \sum_{i=1}^{n_c} x_{ij})$ distribution.

The predictive distribution of the sample y within class c is now a product

Beta-Bernoulli distribution defined as

$$p(y|x_1, ..., x_{n_c}) = \prod_{j=1}^{d} \int_0^1 \theta_{cj}^{y_j}(1-\theta_{cj})^{1-y_j} p(\theta_{cj}|x_1, ..., x_{n_c}) d\theta_{cj}$$

$$= \prod_{j=1}^{d} \int_0^1 \theta_{cj}^{y_j}(1-\theta_{cj})^{1-y_j} \frac{\Gamma(\alpha+\beta+n_c)}{\Gamma(\alpha+\sum_{i=1}^{n_c} x_{ij})\Gamma(\beta+n_c-\sum_{i=1}^{n_c} x_{ij})} \times$$

$$\times \theta_{cj}^{\alpha+\sum_{i=1}^{n_c} x_{ij}-1}(1-\theta_{cj})^{\beta+n_c-\sum_{i=1}^{n_c} x_{ij}-1} d\theta_{cj}$$

$$= \prod_{j=1}^{d} \frac{\Gamma(\alpha+\beta+n_c)}{\Gamma(\alpha+\sum_{i=1}^{n_c} x_{ij})\Gamma(\beta+n_c-\sum_{i=1}^{n_c} x_{ij})} \times$$

$$\times \int_0^1 \theta_{cj}^{y_j+\alpha+\sum_{i=1}^{n_c} x_{ij}-1}(1-\theta_{cj})^{1-y_j+\beta+n_c-\sum_{i=1}^{n_c} x_{ij}-1} d\theta_{cj}$$

$$= \prod_{j=1}^{d} \frac{\Gamma(\alpha+\beta+n_c)}{\Gamma(\alpha+\sum_{i=1}^{n_c} x_{ij})\Gamma(\beta+n_c-\sum_{i=1}^{n_c} x_{ij})} \times$$
$$\times \frac{\Gamma(y_j+\alpha+\sum_{i=1}^{n_c} x_{ij})\Gamma(1-y_j+\beta+n_c-\sum_{i=1}^{n_c} x_{ij})}{\Gamma(\alpha+\beta+n_c+1)},$$

where the last result again follows from the properties of the Beta integral.

Given the above (somewhat breath-taking) derivations, we are finally in the position of presenting the Bayesian predictive classifier formula:

$$\begin{aligned} p(z &= 1|y, x_1, ..., x_{n_1}, x_1, ..., x_{n_2}) = \\ &= \frac{p(y|x_1, ..., x_{n_1})p(z=1)}{p(y|x_1, ..., x_{n_1})p(z=1)+p(y|x_1, ..., x_{n_2})p(z=2)}, \end{aligned} \tag{40}$$

which weighs the predictive likelihoods of observing y under each class against each other.

Typically, the predictive classifier would exhibit more uncertainty about the origin of y since the uncertainty about underlying word frequencies is explicitly taken into account.

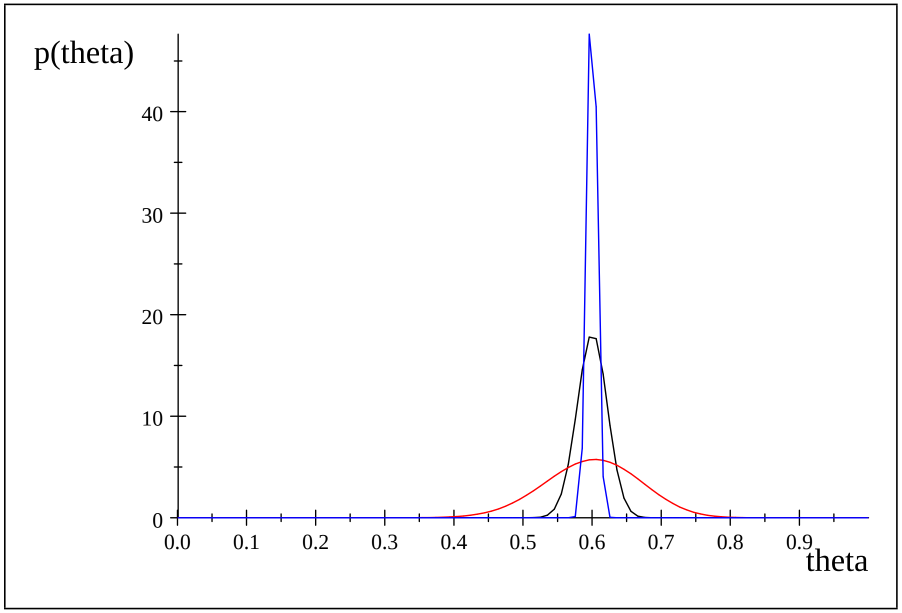
This is particularly important for applications where erroneous classification is associated with large 'costs' or ofther negative consequences, and the applier might be reluctant to classify a sample unless the classification uncertainty is sufficiently negligible.

We now investigate also the limiting behavior of the predictive classifier as a function of the amount of training samples.

Firstly, recall the expectation and variance of the Beta distribution which equal:

$$\begin{aligned} E\theta &= \frac{\alpha}{\alpha+\beta} \\ VAR\theta &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

Examples of Beta density with increasing sum of hyperparameters (50, 500, 5000), with the expectation fixed at 3/5, are shown below.



The above figure provides a clue about the limiting behavior of the predictive distribution

$$p(y|x_1, ..., x_{n_c}) = \prod_{j=1}^{d} \int_0^1 \theta_{cj}^{y_j}(1-\theta_{cj})^{1-y_j} p(\theta_{cj}|x_1, ..., x_{n_c}) d\theta_{cj}. \qquad (41)$$

As $n_c \to \infty$, variance of the posterior distribution diminishes and the distribution becomes increasingly spiky at the maximum likelihood estimate $\hat{\theta}_{cj} = n_c^{-1} \sum_{i=1}^{n_c} x_{ij}$ for each element of $\boldsymbol{\theta}$.

Thus, the predictive distribution will increasingly resemble the distribution $\hat{p}(y|z = c)$ obtained earlier using the point estimates of the word frequencies.

Prior distributions

The most obvious practical difference between classical and Bayesian inference lies in the Bayesian's use of prior information, and careful specification of the prior distribution is of great importance.

In order to improve techniques of subjective probability specification, psychologists have studied how people make probability judgements in practice.

In literature one can find several common errors, and it is clear that in general people do not naturally have good habits of subjective probability assessment.

Probability modelers and others who regularly need to make careful, formal specifications of probabilities should be explicitly trained to do so, and to avoid known pitfalls.

Most often, a statistical analysis of real data demands not the prior probability judgements of a statistician, but of the scientist, experimenter or decision-maker who has collected the data, and who has interest in the inferences. Thus, it is important to have skills in prior elicitation from others.

Consider now specification of prior distributions.

For a discrete θ taking only a few possible values, it is possible to specify a distribution by individually determining the probabilities.

Otherwise, such a procedure is very tedious, and unnecessarily inaccurate, even for discrete θ.

For continuous parameters it is obviously impossible.

Consider instead the effect of specifying a few summaries of Your prior distribution for θ.

For a scalar θ these might include the mean, mode and the standard deviation.

If we add to these Your assessment that the distribution is unimodal, $p(\theta)$ is already rather precisely defined.

Any two distribution satisfying these conditions would generally look somewhat similar.

Individual probabilities would not vary greatly.

More important, we would not expect posterior inferences to be sensitive to the precise choice of prior from among the class satisfying these conditions.

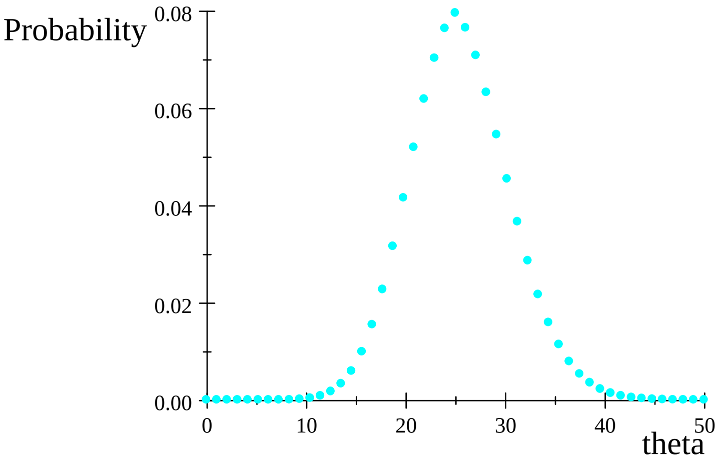
It should be stressed that Bayesian modelling, like any statistical analysis, requires a fair amount of probabilistic knowledge.

Awareness of a wide variety of distributional forms and inference tools facilitates the task of constructing meaningful models for data.

When priors are specified, one should keep in mind the purpose of the model, *i.e.* what are we interested in?

Exactly like the purpose steers the choice of the likelihood, it should affect even the choice of a prior.

Example 12 Poisson prior with a fixed mean. *Figure below shows the Poisson distribution with mean 25. It has mode at 24 and 25 and standard deviation 5. It is hard to construct another distribution with these values and the same general, unimodal shape as the Poisson distribution but with radically different probabilities. Within those constraints any plausible specification of a precise distribution would be expected to produce very similar posterior inferences to those induced by the Poisson prior.*



A simple and effective technique for specifying a prior distribution is to specify an appropriate selection of summaries, and then to adopt any convenient distribution that conforms to these conditions.

Provided enough well-chosen summaries are specified then, for most forms of data, posterior inference should be insensitive to he arbitrary final choice of distribution.

Subjective specification of a distribution in this way is essentially the converse of summarization.

Summarization seeks to express the principal features of a distribution in readily understood terms.

The specification process takes information expressed in those same readily understood terms and tries to find an appropriate distribution.

In either direction, summaries are the natural way of conveying information about θ.

For instance, to be told that the posterior mean of θ is 1.5 conveys a useful point estimate of θ.

Conversely, one may make a prior estimate of θ by specifying the prior mean $E(\theta)$.

However, it is important to understand the sense in which the mean represents an estimate, or a location summary, of a distribution.

To be told that the mode is 1.5 conveys slightly different information, and to be told that the median is 1.5, has yet more different information content.

A tricky question is how to select a distribution satisfying the specified prior summaries.

Given that at this point it is acceptable to use any distribution that agrees with those summaries, the choice is arbitrary.

It is usually made on the grounds of convenience, in one of two senses.

It may just be the first distribution that springs to mind to fit the summaries, or a distribution of a general form for which it is easy to fit them.

For instance, if prior mean and variance of a scalar θ are given, then an easy choice is the normal distribution with the given mean and variance.

Or if θ is necessarily positive, we could instead fit a gamma distribution having those moments.

Although prior distributions are often chosen in the above described manner, there is also a more important sense in which a choice of a prior can be convenient.

Suppose that data are to be observed with distribution $p(x|\theta)$.

A family $\mathcal{F}$ of prior distributions for θ is said to be *closed under sampling* (or sometimes conjugate) from $p(x|\theta)$ if for every prior distribution $p(\theta) \in \mathcal{F}$, the posterior distribution $p(\theta|x)$ is also in $\mathcal{F}$.

We encountered this feature in the thumbtack example, where Beta distribution was used as a prior for the probability of observing a toss with point up.

The conjugate priors are often also useful for parameters at different levels of hierarchical models.

We will examine closer choice of priors in case study from a clustering context and the consequences of different choices (see the separate pdf document).

Selection of prior distributions by formal rules

Earlier we have seen how the Bayesian paradigm provides a normative tool for probabilistic data analysis by certain rules that are available for coherent quantitative learning given empirical observations.

The problem, however, is that many real-world phenomena we are interested in are of such complexity that the specification of one's subjective beliefs in a reasonable probabilistic form is a daunting task.

Also, use of the numerical methods necessary to obtain the sought answers requires typically a fair amount of expertise.

Another issue that is often taken up in the "choice of priors" debate, is that of representation of ignorance in order not to "bias" the information the data has to say, for instance, about a parameter θ.

In particular, such "dream of objectivism" has led to the development of mathematical rules which for specified problems aim to produce an automated answer to the question: What is Your prior opinion?

The result, of course, is that anyone applying the same rule, will express the same prior opinion for the problem at hand.

This, combined with the idea that the answers provided by the rules should be minimally informative or as vague as possible, leans on the "dream of objectivism".

The less information you provide for the problem, the more the data have to say, and the answers then produced could be understood as approximate consensus among possible modelers who have only vague prior opinions.

An extensive review of different rules to produce automated prior opinions (we call them here reference priors) is given in Kass and Wasserman (1996).

Two interpretations of reference priors are of main importance.

The first interpretation asserts that reference priors are formal representations of ignorance.

The second asserts that there is no objective, unique prior that represents ignorance.

Instead, reference priors are chosen by public agreement, much like units of weight and length.

In this interpretation, reference priors are akin to a default option in a computer package.

We fall back to the default when there is insufficient information to otherwise define the prior.

In principle, we could construct a systematic catalogue of reference priors for a variety of models.

The priors in catalogue do not represent ignorance, but are useful in problems where it is too difficult to elicit an appropriate subjective prior.

A data modeler may feel that the reference prior is, for all practical purposes, a good approximation to any reasonable subjective prior for that problem.

The first mentioned interpretation of reference priors was earlier a prominent view in Bayesian probability modeling.

However, currently the public opinion is in favor of the second interpretation, and it is difficult to imagine anyone claiming that a particular prior can logically be defended as being truly noninformative.

Instead, research in this field is focused on particular prior deriving procedures, to see if any of them have advantages over the others in some practical sense.

If the parameter space is finite, then *Laplace's rule*, or the *principle of insufficient reason*, is to use a uniform prior that assigns equal probability to each point in the parameter space.

Use of uniform probabilities on finite sets dates back to the origins of probability in gambling problems.

The terminology comes from references by Laplace to a lack of sufficient reason for assuming nonuniform probabilities.

Example 13 Application of Laplace's rule. *Consider a scenario where You have to specify probabilistic beliefs about how much the lecturer of this course has money in his pockets at today's lecture. Clearly, in a given currency, say EUR, this sum must be finite. However, for illustrative purposes we may also consider the possibility of accepting an infinite amount as an answer. If the lecturer's pockets are not bulging extensively, You might consider that no more than one hundred 500 EUR bills could be fitted in the pockets without a notice. If You then applied Laplace's rule, the prior would specify the same degree of belief to finding after inspection a single one cent coin as finding 50000 EUR. I hope that You certainly have a sufficient reason to be suspicious about such a prior.*

Example 14 Application of Laplace's rule continued. *Now, consider the case where the amount of money is allowed to lie in the interval* $[0, \infty)$. *If we now apply a generalization of Laplace's rule, a prior which is constant over this interval is obtained. It is clear that the prior is* **improper** *in the sense that it does not specify a probability density. However, similar priors have been suggested and used in the statistical literature. A remarkable property of such a prior is, for instance, that You assign ultimately much more belief in the event that the amount of money exceeds 50000 EUR than in the event that it is at most 50000 EUR.*

When applying rules such as Laplace's to obtain a prior distribution one needs to be aware of the pitfalls that might be encountered.

For instance, a refinement of the parameter space leads to different prior beliefs.

Let $\Theta = \{\theta_1, \theta_2\}$, where θ_1 denotes the event that there is life in orbit about the star Sirius and θ_2 denotes the event that there is not.

Clearly, Laplace's rule gives $P(\theta_1) = P(\theta_2) = 1/2$.

Let now $\Omega = \{\omega_1, \omega_2, \omega_3\}$, where ω_1 denotes the event that there is life around Sirius, ω_2 denotes that there are planets but no life, and ω_3 denotes that there are no planets.

Laplace's rule gives $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$, so that the probability of life has fallen considerably.

To avoid this type of problems, one should use scientific judgement to choose a particular level of refinement that is meaningful for the problem at hand.

Another problem appears when applying a uniform prior to a parameter which is continuous, namely that the prior is not invariant under transformation.

If we start with a Uniform(0,1) distribution for ϕ, then $\theta = \log \phi$ will not have a uniform distribution.

To avoid such paradoxes under Laplace's rule, we need to determine a privileged parametrization or change our prior opinion if we change the parametrization.

The uniform prior beliefs are also be obtained when applying the *maximum entropy principle* in the case with finite parameter space, where no further constraints are imposed on the parameters.

So called Shannon entropy of a distribution $p(\theta)$ is defined as

$$h_p = -\sum p(\theta) \log p(\theta) \tag{42}$$

and it is a central concept in information theory.

For any discrete event space the entropy is maximized for the uniform distribution.

In the continuous case, however, one needs to choose a suitable base measure dP according to which the entropy is defined.

This, in turn is almost as difficult as the choice of a prior so that the maximum entropy principle has restricted application in continuous problems.

A widely considered approach to deriving reference priors is the Jeffreys' method for location-scale problems and its several later ramifications.

Let $\mathbf{I}(\boldsymbol{\theta})$ denote the *Fisher information matrix*, defined under specific regularity conditions as

$$\mathbf{I}(\boldsymbol{\theta})_{ij} = -E\left(\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right) \tag{43}$$

where l is the log-likelihood function.

Notice that the above expectation is over the sample space.

If there are location parameters in the model, say labeled by $\mu_1, ..., \mu_k$, and some additional parameters (possibly including scale parameters) $\boldsymbol{\theta}$, then the prior Jeffreys derived becomes

$$p(\mu_1, ..., \mu_k, \boldsymbol{\theta}) \propto \mathsf{det}(\mathbf{I}(\boldsymbol{\theta}))^{1/2}, \tag{44}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is calculated holding the location parameters fixed.

Notice that due to the invariance property of the Fisher information, the above prior is invariant under one-to-one transformations of the parameters in $\boldsymbol{\theta}$.

A general drawback of reference priors is that they typically do not correspond to probability measures on the parameter space $\boldsymbol{\Theta}$, but simply to functions (often called improper priors), say $g(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$.

So in the strict sense, any analysis based on such prior opinions falls automatically outside the Bayesian paradigm.

Nevertheless, many scientists have accepted the use of improper priors as an approximation to a strict Bayesian analysis, and indeed, in many cases the approximations are sensible.

For inference about a fixed-dimensional parameter $\boldsymbol{\theta}$ the introduction of an improper prior $g(\boldsymbol{\theta})$ need not be as prohibitive as might be expected at a first glance. In such cases the prior opinion is often expressed as

$$p(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta}) \tag{45}$$

where the proportionality means that the prior distribution is understood as proportional to a real-valued function $g(\boldsymbol{\theta})$ for which $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \neq 1$, but which satisfies

$$\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty \tag{46}$$

Under this assumption the "posterior"

$$\frac{p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{47}$$

will be a well-defined probability measure since the unknown "proportionality constant" cancels in the above formula.

In many cases the "Bayesian inference" (proceeding with the paradigm as if

$g(\boldsymbol{\theta})$ was a real prior) based on such procedure can be shown to have acceptable properties in the statistical sense.

Intuitively, this can be understood since $g(\boldsymbol{\theta})$ is typically a close approximation (or a limit) to a real, vague prior density.

However, there are also numerous cases where the performance is unacceptable, so caution needs to be taken whenever reference priors are used (in fact the same argument applies to proper priors as well).

We will return to the issue of choosing priors automatically in a later section.

What is good inference?

Classical inference theory is very concerned with constructing *good* inference rules.

For instance, since any function of the data can formally serve as a classical estimator of a parameter, it is necessary to identify criteria for comparing estimators, and for saying that one estimator is somehow better than another.

The primary concern of basic Bayesian inference is entirely different, since the objective is to extract information concerning the model parameters from the posterior, and to present it helpfully via effective use of summaries.

This procedure should obey rules of good communication.

Summaries should be chosen to convey clearly and succintly all the features of interest.

In this framework it is not possible to construct a formal mathematical structure to measure how good inference is.

We cannot even say what we mean by "interesting features of the posterior distribution".

Interest is very dependent on context.

We can give examples of the sorts of features that are likely to be of interest, develop a strategy to identify them, and construct good summaries to display them, but interest often resides in the unusual or the unexpected.

When we have a complex, multidimensional posterior distribution, we can never be sure that we have summarized it exhaustively.

This is one aspect of the statistician's work that relies heavily on experience.

In Bayesian terms, therefore, a good inference is one which contributes effectively to appreciating the information conveyed by the posterior distribution.

Professor Daniel Thorburn, at the Department of Statistics, Stockholm University, once defined Bayesian inference like a sharp knife. Such a device is extremely helpful when one wishes to carve a nice boat out of a piece of wood. On the other hand, the same device in less careful hands may result in a deep, bleeding cut. Certain other means of inference he categorized as a plastic knife instead. Such a device is not of much use in carving, but even a careless person can use it without harming himself.

Hierarchical models and partial exchangeability

In the previous sections we have investigated various kinds of justification for modeling a sequence of random quantities as a random sample from a parametric family with density $p(x|\theta)$, together with a prior distribution $Q(\theta)$ for θ.

However, in order to concentrate on the basic conceptual issues, we have thus far restricted attention mostly to the case of a single sequence of random quantities labeled by a single index, and unrelated to other random quantities.

Clearly, in many areas of modeling applications the situation will be more complicated than this, and we need to elaborate the basic formulation.

A case study with occurrence patterns of cancers illustrates the usefulness of a hierarchical modeling approach (see the separate example pdfs).

Suppose, for example, that several treatments are administered in a clinical trial.

From each treatment group we will make some observations.

It may be plausible to model the observations within each treatment group as exchangeable, but it would seem strange to model all observations as exchangeable.

For each treatment group, we might develop a parametric model as we have done earlier.

A *hierarchical model* for this example involves treating the set of parameters corresponding to the different treatment groups as a sample from another population.

Prior to seeing any observations, we can model the parameters as exchangeable.

This would mean that we could introduce another set of parameters to model their joint distribution.

These second-level parameters are typically called *hyperparameters*.

Example 15 Normal response in different treatment groups. *Suppose that there are k treatment groups. Let x_{ij} stand for the observed response of subject j in treatment group i. We might invent parameters $M_1, ..., M_k$ and model the responses x_{ij} as conditionally independent random normal quantities $N(\mu_i, 1)$ given $(M_1, ..., M_k) = (\mu_1, ..., \mu_k)$. We could then model $M_1, ..., M_k$ as a priori exchangeable with distribution $N(\theta, 1)$ given θ. Here, θ is a hyperparameter for which we need to specify a belief distribution. Notice that here we have only one θ regardless of the value of k.*

The intuitive concept of how hierarchical models work is the following.

Suppose that the data comprise several groups, each of which we consider to be a collection of exchangeable random quantities.

From the data in each group, we obtain direct information about the corresponding parameters.

Thinking of the hyperparameters as known for the time being, we then update the distributions of the parameters using the data, to get posterior distributions for the parameters via Bayes' theorem.

Future data (in each group) are still exchangeable with the same conditional distributions given the parameters, but the distributions of the parameters have changed. In fact, the distribution of each parameter (given the hyperparameters) has now been updated using only the data from its corresponding group.

Bayes' theorem can now be used again to find the posterior distribution of the hyperparameters given the data.

The marginal posterior of the parameters given the data is found by integrating the hyperparameters out of the joint posterior of the parameters and hyperparameters.

This is how the data from all groups combine to provide information about all of the parameters, not just the ones corresponding to their own group.

It is the common dependence of all parameters on the hyperparameters that allows us to make use of common information in updating the distribution of all parameters.

In theory, the updating of information can be performed as follows.

Let future observations again be denoted by $\mathbf{y}$ and the current observations by $\mathbf{z}$.

Let $\boldsymbol{\theta}$ be the parameters and $\boldsymbol{\psi}$ the hyperparameters.

The joint distribution of $(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\psi})$ is intuitively specified by recursive conditioning as

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi}) \tag{48}$$

The posterior density of the parameters given the hyperparameters is

$$p(\boldsymbol{\theta}|\mathbf{z}, \boldsymbol{\psi}) = \frac{p(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})}{p(\mathbf{z}|\boldsymbol{\psi})} \tag{49}$$

where the density of the data given the hyperparameters alone is

$$p(\mathbf{z}|\boldsymbol{\psi}) = \int p(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\boldsymbol{\psi})d\boldsymbol{\theta}. \tag{50}$$

The marginal posterior of the parameters can be found from

$$p(\boldsymbol{\theta}|\mathbf{z}) = \int p(\boldsymbol{\theta}|\mathbf{z}, \boldsymbol{\psi})p(\boldsymbol{\psi}|\mathbf{z})d\boldsymbol{\psi} \tag{51}$$

where the posterior density of $\boldsymbol{\psi}$ given $\mathbf{z}$ is

$$p(\boldsymbol{\psi}|\mathbf{z}) = \frac{p(\mathbf{z}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\mathbf{z})} \tag{52}$$

and the marginal density of the data is

$$p(\mathbf{z}) = \int p(\mathbf{z}|\boldsymbol{\psi})p(\boldsymbol{\psi})d\boldsymbol{\psi} \tag{53}$$

Finally, under the assumption of conditional independence between the future $\mathbf{y}$ and current observation given the parameters and hyperparameters, the predictive distribution of $\mathbf{y}$ can be written as

$$p(\mathbf{y}|\mathbf{z}) = \int\int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi})p(\boldsymbol{\theta}|\mathbf{z}, \boldsymbol{\psi})p(\boldsymbol{\psi}|\mathbf{z})d\boldsymbol{\theta}d\boldsymbol{\psi}. \tag{54}$$

As we saw in Example 11 where tosses were made with different thumbtacks, exchangeability in its basic formulation may not be a reasonable assumption in situations with built-in heterogeneity in the observation scheme.

The generalization of exchangeability to account for such heterogeneity is a rather deep issue, and there is no universal definition we could rely on.

Hence, consideration of partial exchangeability is much dependent on the particular modeling situation under investigation.

Definition 5 Unrestricted exchangeability for binary sequences. *Let* $\mathbf{x}_i(n_i)$ *denote a vector of binary random quantities* $x_{i1}, ..., x_{in_i}, i = 1, ..., m$. *Sequences of binary random quantities* $x_{i1}, x_{i2}, ...,$ $i = 1, ..., m$, *are said to be unrestrictedly exchangeable if each sequence is infinitely exchangeable and, in addition, for all* $n_i \leq N_i$, $i = 1, ..., m$,

$$p(\mathbf{x}_1(n_1), ..., \mathbf{x}_m(n_m)|y_1(N_1), ..., y_m(N_m)) = \prod_{i=1}^{m} p(\mathbf{x}_i(n_i)|y_i(N_i)) \quad (55)$$

where $y_i(N_i) = x_{i1} + \cdots + x_{iN_i}, i = 1, ..., m$.

In addition to the exchangeability of the individual sequences, this definition encapsulates the judgement that, given the total number of successes in the first N_i observations from the ith sequence, $i = 1, ..., m$, *only the total for the* i*th sequence is relevant* when it comes to beliefs about the outcomes of any subset of n_i of the N_i observations from that sequence.

The unrestricted exchangeability implies that

$$\begin{aligned} & p(x_{11}, ..., x_{1n_1}, ..., x_{m1}, ..., x_{mn_m}) && (56) \\ = \; & p(x_{1\pi_1(1)}, ..., x_{1\pi_1(n_1)}, ..., x_{m\pi_m(1)}, ..., x_{m\pi_m(n_m)}) && (57) \end{aligned}$$

for any unrestricted choice of permutations π_i of $\{1, ..., n_i\}, i = 1, ..., m$.

For example, given 15 deaths in the first 100 patients receiving Drug 1 ($N_1 = 100, y_1(N_1) = 15$) and 20 deaths in the first 80 patients receiving Drug 2 ($N_1 = 80, y_1(N_1) = 20$), we would typically judge the latter information to be irrelevant to any assessment of the probability that the first tree patients receiving Drug 1 survived and the fourth one died ($x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{14} = 1$). Given the definition of unrestricted exchangeability, we may establish a generalization of the earlier stated representation theorem.

Proposition 6 Representation theorem for several sequences of binary random quantities. *If* $x_{i1}, x_{i2}, ...,$ $i = 1, ..., m,$ *are unrestrictedly infinitely exchangeable binary random sequences with joint probability measure* P*, there exists a distribution function* Q *such that*

$$p(\mathbf{x}_1(n_1), ..., \mathbf{x}_m(n_m)) = \int_{[0,1]^m} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \theta_i^{x_{ij}} (1 - \theta_i)^{1-x_{ij}} dQ(\boldsymbol{\theta}) \qquad (58)$$

where $y_i(n_i) = x_{i1} + \cdots + x_{in_i}, i = 1, ..., m,$ *and*

$$Q(\boldsymbol{\theta}) = \lim_{all\ n_i \to \infty} P\left[\left(\frac{y_1(n_1)}{n_1} \leq \theta_1\right) \cap \cdots \cap \left(\frac{y_m(n_m)}{n_m} \leq \theta_m\right)\right]. \qquad (59)$$

Let us investigate, for simplicity, a modeling situation under the above scenario with $m = 2$.

Our belief model will be completed by the specification of $Q(\theta_1, \theta_2)$ whose detailed form will, of course, depend on the particular beliefs considered appropriate.

Two examples of beliefs are as follows:

- Knowledge of the behavior of one of the sequences would not change beliefs about outcomes in the other sequence, so that we have the independent form of prior specification $Q(\theta_1, \theta_2) = Q(\theta_1)Q(\theta_2)$.
- The limiting relative frequency for the second sequence will be necessary greater than that for the first sequence, so that $Q(\theta_1, \theta_2)$ is zero outside the range $0 \leq \theta_1 < \theta_2 \leq 1$.

Example 16 Heterogeneous thumbtack example continued. *Suppose we make tosses with $k = 2$ thumbtacks made of different materials. We can model the parameters $\theta_i, i = 1, ..., k$, as exchangeable Beta($\mu\lambda, (1 - \mu)\lambda$) random quantities. Here μ is like the average probability of observing a toss with point up and λ is like a measure of similarity, since the larger λ is, the more similar will θ_i's be. The posterior distribution of θ_i given ($\mu, \lambda, \sum_{j=1}^{n_i} x_{ij}$) is Beta($\mu\lambda+\sum_{j=1}^{n_i} x_{ij}, (1-\mu)\lambda+n_i-\sum_{j=1}^{n_i} x_{ij}$). Correspondingly, the posterior density of (μ, λ) is proportional to*

$$p(\mu, \lambda)\frac{\Gamma(\lambda)^k}{\Gamma(\mu\lambda)^k\Gamma((1-\mu)\lambda)^k} \times \tag{60}$$

$$\times \prod_{i=1}^{k} \frac{\Gamma(\mu\lambda + \sum_{j=1}^{n_i} x_{ij})\Gamma((1-\mu)\lambda + n_i - \sum_{j=1}^{n_i} x_{ij})}{\Gamma(\lambda + n_i)}. \tag{61}$$

Basic Bayesian inference procedures

We have earlier considered representation and revision of beliefs as the basis of empirical learning.

Here we shall investigate some simple examples.

Example 17 Single observation from a normal distribution. *Let x have a normal distribution $N(\theta, v)$ with unknown mean θ and known variance v, and let the prior distribution for θ be $N(m, w)$. Let the precision parameters be $\lambda_0 = 1/v$ and $\lambda_1 = 1/w$. Then,*

$$\begin{aligned} p(x|\theta, v) &= \frac{1}{\sqrt{2\pi v}} \exp(-\frac{1}{2v}(x-\theta)^2) \qquad (62) \\ p(\theta|m, w) &= \frac{1}{\sqrt{2\pi w}} \exp(-\frac{1}{2w}(\theta-m)^2). \end{aligned}$$

By multiplying together the prior and the likelihood, and expanding the squares, we get the exponential

$$\exp\left(-\frac{1}{2}\lambda_0(x^2 - 2x\theta + \theta^2) - \frac{1}{2}\lambda_1(\theta^2 - 2\theta m + m^2)\right). \qquad (63)$$

The exponential can be further written as

$$
\begin{aligned}
& -\frac{1}{2}\lambda_0 x^2 + \lambda_0 x\theta - \frac{1}{2}\lambda_0\theta^2 - \frac{1}{2}\lambda_1\theta^2 + \lambda_1\theta m - \frac{1}{2}\lambda_1 m^2 \qquad (64)\\
= \; & -\frac{1}{2}(\lambda_0 + \lambda_1)\theta^2 + \theta(\lambda_0 x + \lambda_1 m) - \frac{1}{2}(\lambda_0 x^2 + \lambda_1 m^2)\\
= \; & -\frac{1}{2}(\lambda_0 + \lambda_1)\left(\theta - 2\theta\frac{\lambda_0 x + \lambda_1 m}{\lambda_0 + \lambda_1} + \left(\frac{\lambda_0 x + \lambda_1 m}{\lambda_0 + \lambda_1}\right)^2\right) + c
\end{aligned}
$$

where c does not depend on θ. Since the constants cancel in

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}, \qquad (65)$$

the posterior is recognized as the density function of the normal distribution

$$N\left(\frac{\lambda_0 x + \lambda_1 m}{\lambda_0 + \lambda_1}, \lambda_0 + \lambda_1\right), \qquad (66)$$

where the mean is a weighted average of prior mean m and observation x.

Therefore the posterior mean (as well as mode and median) is a compromise between the prior information and the sample information.

We see also that each source of information is weighted proportionately to its precision.

Consequently, the posterior mean will lie closer to whichever source has the stronger information.

If, for instance, prior information is very weak, expressed by λ_1 being close to zero, then the posterior mean will be close to x.

The posterior precision is the sum of the prior and data precisions, reflecting the combination of information from the two sources.

The posterior information is stronger than either source of information alone.

Example 18 Several observations from a normal distribution. *In the previous example we had only a single observation available for making inference about the mean of the distribution. However, typically, we would utilize several observations. Let* $x_1, ..., x_n$ *be conditionally independent observations from a normal distribution* $N(\theta, 1)$ *with unknown mean* θ *and known variance* 1. *Suppose the prior distribution for* θ *is again* $N(m, w)$, *i.e. the precision parameter is* $\lambda = 1/w$. *The likelihood function can be written as*

$$p(\mathbf{x}|\theta) = (2\pi)^{-n/2} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right), \qquad (67)$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ *is the sample mean. Multiplying the likelihood and prior together and simplifying yields the following expression for the numerator* *of the posterior formula,*

$$\exp\left(-\frac{n+\lambda}{2}\left(\theta-\frac{\lambda m+n\bar{x}}{\lambda+n}\right)^2\right), \qquad (68)$$

thus, the posterior is $N(\frac{\lambda m+n\bar{x}}{\lambda+n}, 1/(\lambda+n))$*. We see that the posterior variance decreases (i.e. the precision increases) as the sample size increases, and similarly that the dependence on the prior mean decreases as well.*

Example 19 Predictive distribution of a future observation. *Let us continue analysis of the previous example by considering the predictive density of a future observation* x_{n+1}

$$
\begin{aligned}
& p(x_{n+1}|\mathbf{x}) && (69)\\
= \; & \int p(x_{n+1}|\theta)p(\theta|\mathbf{x})d\theta && (70)\\
= \; & \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right) \frac{\sqrt{n+\lambda}}{\sqrt{2\pi}} \exp\left(-\frac{n+\lambda}{2}\left(\theta - \frac{\lambda m + n\bar{x}}{\lambda + n}\right)^2\right) d\theta \\
= \; & \frac{\sqrt{n+\lambda}}{\sqrt{2\pi(n+\lambda+1)}} \exp\left(-\frac{n+\lambda}{2(n+\lambda+1)}\left(y - \frac{\lambda m + n\bar{x}}{\lambda + n}\right)^2\right),
\end{aligned}
$$

which is the density of the normal distribution $N(\frac{\lambda m + n\bar{x}}{\lambda+n}, 1 + 1/(\lambda+n))$. *Thus, we see that the excess uncertainty in the predictive distribution, which is due to the "estimation" of the unknown parameter* θ*, vanishes as the sample size tends to infinity. This procedure is in perfect harmony with intuition about* *how information is gathered and utilized.*

Example 20 Observations from a Poisson distribution. *Let x have the Poisson distribution with unknown mean θ,*

$$p(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}, \tag{71}$$

and suppose that the prior density has the Gamma(α, β) form

$$p(\theta) = \frac{\alpha^\beta \theta^{\beta-1}}{\Gamma(\beta)} e^{-\alpha\theta}, \theta > 0. \tag{72}$$

by combining the prior and the likelihood we enter into the Gamma($\alpha+1, \beta+x$) posterior. When the likelihood comprises n observations $x_1, ..., x_n$

$$p(x|\theta) = \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\theta},$$

use of the same prior as above, gives us the Gamma($\alpha + n, \beta + \sum_{i=1}^n x_i$) posterior. The mean of this distribution equals $\beta + \sum_{i=1}^n x_i/(\alpha + n)$ and the *variance* $\beta + \sum_{i=1}^{n} x_i/(\alpha + n)^2$.

Example 21 Bayesian estimation of Shannon entropy. *We now consider a considerably more complicated inference situation than encountered in the previous examples, taken from Yuan and Kesavan (1997). Recall from the previous chapter the concept of the entropy for a discrete random quantity x taking values conveniently labeled by a finite set of integers $\{1, ..., s\}$ associated with a probability distribution $\mathbf{p}$ =$(p_1, ..., p_s)$ satisfying $p_i > 0, i = 1, ..., s$, and $\sum_{i=1}^{s} p_i = 1$. The entropy is defined as*

$$h = -\sum_{i=1}^{s} p_i \log p_i. \tag{73}$$

Here we use the natural logarithm in the definition of entropy, however, other bases are also often used in the literature. If the true distribution is known, then the calculation of the entropy is straightforward. In practice, however, we often have to estimate h from data under no or vague knowledge about the underlying probability distribution $\mathbf{p}$. Suppose we have frequency data generated from a *multinomial distribution* $\mathbf{p}$, *leading to the likelihood*

$$\binom{n}{n_1 \cdots n_s} p_1^{n_1} \cdots p_s^{n_s} \tag{74}$$

where $n = \sum_{i=1}^{s} n_i$ *and* $\binom{n}{n_1 \cdots n_s}$ *is the multinomial coefficient. We recall from earlier that the maximum likelihood estimate* $\hat{p}_i$ *of* p_i *is provided by the observed relative frequency* $n_i/n, i = 1, ..., s$. *Apparently, this procedure leads to the entropy estimate*

$$h_n = -\sum_{i=1}^{s} \hat{p}_i \log \hat{p}_i. \tag{75}$$

While the above estimate may be deemed satisfactory for large n *relative to* s, *its properties could be improved upon when the converse is true. From the definition of entropy we see that the values of* x *having zero observed frequencies make no contribution to the estimate* h_n.

Assume we have a prior guess about the unknown distribution $\mathbf{p}$*, say* $\boldsymbol{\pi} = (\pi_1, ..., \pi_s)$*, with* $\sum_{i=1}^{s} \pi_i = 1, \pi_i > 0$*. We could now use the Dirichlet* $D(\alpha\pi_1, ..., \alpha\pi_s)$ *distribution to describe our prior beliefs, where the parameter* α *is a measure of our confidence about our guess. A larger value of* α *implies more concentration of the prior around* $(\pi_1, ..., \pi_s)$*. If we do not have any prior knowledge, a uniform prior* $D(1, ..., 1)$ *could be used.*

Under the above Dirichlet prior we get an explicit expression for the posterior mean of the entropy, which equals

$$h_B = -\sum_{i=1}^{s} \frac{\alpha\pi_i + n_i}{\alpha + n} \left[\psi(\alpha\pi_i + n_i + 1) - \psi(\alpha + n + 1)\right], \qquad (76)$$

where $\psi(t) = \Gamma'(t)/\Gamma(t)$ *is the digamma function. When* α *is large compared with* n*,* h_B *is mainly determined by the prior, and consequently, the contribution of the data is small. With the increase of* n*, the behavior of* h_B *is as that of*

h_n. *When the prior is uniform we get the expression*

$$h_{B_0} = -\sum_{i=1}^{s} \frac{1+n_i}{s+n} \left[\psi(n_i+2) - \psi(s+n+1)\right]. \qquad (77)$$

Robustness and sensitivity

A major question in any application of Bayesian methods is the extent to which the inferences are sensitive to possible mis-specification of the prior distribution or the likelihood.

And if different specifications lead to different inferences, can we determine which is the 'correct' or 'better' specification?

In most real applications of probability modeling, it has to acknowledged that both prior distribution and likelihood have only been specified as more or less convenient approximations to whatever the investigator's true belief might be.

If the inferences from the Bayesian analysis are to be trusted, it is important to determine that they are not sensitive to such variations of prior and likelihood as might also be consistent with the investigator's stated beliefs.

In arriving at a particular parametric model specification, by means of whatever combination of formal and pragmatic judgements have been deemed appropriate, a number of simplifying assumptions will necessarily have been made (either consciously or unconsciously).

Therefore, it would always be prudent to try to review the judgements that have been made. For instance, one might ask questions like:

- Is it reasonable to assume that all the observables form a "homogeneous sample", or might a few of them be "aberrant" in some sense?
- Is it reasonable to apply the modeling assumptions to the observables on their original scale of measurement, or should the scale be transformed to logarithms, reciprocals, or whatever?
- When considering temporally or spatially related observables, is it reasonable to have made a particular conditional independence assumption, or

should some form of dependence be taken into account?

- If some, but not all, potential covariates have been included in the model, is it reasonable to have excluded the others?

Several procedures for checking robustness are discussed in the literature. However, their suitability depends firmly on the modeling situation at hand, which means that a general discussion is difficult at the level of the current material. Instead, this topic will be examined through computer examples.

Model comparison, Part I

The choice of an appropriate structure to represent features observed in data is an inherent part of data analysis.

In statistical modeling this corresponds to the choice of an appropriate probability model and necessitates formulation of the model structure and often also estimation of its dimension.

The question of *how* one should compare probability models is to a large extent a philosophical issue.

In broad terms a widely accepted Occam's razor principle says that unnecessary parts should be eliminated from a scientific theory, *i.e.* the parts which cannot be empirically verified.

In the current context it is quite natural to restrict ourselves to the empirical verification, although other forms, such as the use of logic, can be considered in general.

We shall see that the Occam's razor principle is automatically built in the subjectivist's approach.

In reality, in general, there always exists a discrepancy between models and observations.

Such a discrepancy might already arise from the consideration of the accuracy of some measurements that are made and give rise to our data.

Second, since all data analysis is performed by computers where any numbers are represented by finite precision binary digits, it is natural to view models involving densities for continuous variables as approximations for the data.

Nevertheless, to give us guidance in the construction of tools for model comparison, it is sometimes useful to imagine a "computer game scenario", where the observations are generated from a model which belongs to a class of models known to us.

The uncertainty in this situation arises from the fact that we don't know which of the models can be seen as responsible for the data.

This imaginary construct enables us to distinguish good ideas (model comparison strategies) from the less good ones in the ideal world.

What one really hopes then, is that the solutions found to be good in the computer game scenario, would continue to be good in the real world if our models are sufficiently good descriptions of the regularities involved in the phenomenon under investigation.

Conversely, if a model comparison strategy turns out to be a poor one in the ideal world, we expect it to be a poor one even in the real world.

From the Bayesian point of view, one could easily say that models exist in our heads, representing subjective beliefs about some phenomenon.

They are abstract constructs aimed to give perceivable structures to complicated real-world phenomena that can be communicated with others.

The typical statistical interpretation of models as data generating machines is not very realistic or elaborate in this respect (remember the earlier quote of Rissanen).

At least three broad goals can be distinguished in statistical analysis:

- Estimation of unknown parameters
- Model comparison
- Prediction.

We will see that for some approaches these goals can in fact be united.

Model comparison, Part II

Given the fundamental concepts from the previous chapters, let us now consider the model comparison issue in the Bayesian framework.

To proceed as concretely as possibly, assume that all elements in our class of belief models $\mathcal{I}$ are such that the joint density of the observations may be described in terms of a finite-dimensional parameter.

Thus, the predictive distributions (also called marginal likelihoods) for the alternative models are described by

$$p_i(\mathbf{x}) = p(\mathbf{x}|M_i) = \int p_i(\mathbf{x}|\boldsymbol{\theta}_i)p_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i,\ i \in \mathcal{I} \tag{78}$$

Notice, that each element in I constitutes according to our beliefs a possible predictive model for the data.

In order to proceed coherently in the Bayesian framework we then have to build an overall belief model for $\mathbf{x}$ by assigning weights to the different alternatives, so that our subjective beliefs are represented probabilistically.

The overall model takes the form

$$p(\mathbf{x}) = \sum_{i \in \mathcal{I}} P(M_i) p(\mathbf{x}|M_i) \tag{79}$$

where $P(M_i)$ are the weights of the individual belief models such that

$$\sum_{i \in \mathcal{I}} P(M_i) = 1 \tag{80}$$

.

In the literature there has been a considerable amount of discussion about the interpretation of the weights $P(M_i)$.

On one hand, they may be considered as *a priori* probabilities that the corresponding models are "true".

On the other hand, they can simply be regarded as means of representing the degree of dominance (or perhaps functions of odds) of a subjective belief over another.

We first investigate the case where the only action to be taken is a choice of a model $M_i, i \in \mathcal{I}$.

Let ω be an unknown of interest, such that the utility function for our decision problem has the form $u(M_i, \omega)$.

Using the decision theoretic approach we know that the optimal decision is to choose the model M^* which maximizes the expected utility according to

$$\bar{u}(M^*|\mathbf{x}) = \sup_{i \in \mathcal{I}} \bar{u}(M_i|\mathbf{x}) \tag{81}$$

where

$$\bar{u}(M_i|\mathbf{x}) = \int u(M_i, \omega) p(\omega|\mathbf{x}) d\omega, \ i \in \mathcal{I} \tag{82}$$

where $p(\omega|\mathbf{x})$ represents the beliefs about ω having observed $\mathbf{x}$. These have further the form

$$p(\omega|\mathbf{x}) = \sum_{i \in \mathcal{I}} p_i(\omega|M_i, \mathbf{x}) P(M_i|\mathbf{x}) \tag{83}$$

where

$$P(M_i|\mathbf{x}) = \frac{P(M_i)p(\mathbf{x}|M_i)}{\sum_{i\in\mathcal{I}} P(M_i)p(\mathbf{x}|M_i)} \tag{84}$$

is the posterior predictive weight or subjective posterior probability of the individual model being the "true" model.

Notice that there is certainly nothing wrong in the latter definition if one restricts the interpretation to an observer's narrow (or perhaps naive) perception of the world.

If we let the above unknown of interest ω be simply the "true" model among those in $\mathcal{I}$, the decision problem is concretized as follows.

A natural utility function takes the form (0-1 loss)

$$u(M_i, \omega) = \begin{cases} 1 \text{ if } \omega = M_i \\ 0 \text{ if } \omega \neq M_i \end{cases} \tag{85}$$

It then follows that

$$p_i(\omega | M_i, \mathbf{x}) = \begin{cases} 1 \text{ if } \omega = M_i \\ 0 \text{ if } \omega \neq M_i \end{cases} \tag{86}$$

and

$$p(\omega | \mathbf{x}) = \begin{cases} P(M_i | \mathbf{x}), & \text{if } \omega = M_i \\ 0, & \text{if } \omega \neq M_i \end{cases} \tag{87}$$

The expected utility of the choice M_i is

$$\begin{aligned} \bar{u}(M_i|\mathbf{x}) &= \int u(M_i,\omega)p(\omega|\mathbf{x})d\omega \qquad (88) \\ &= P(M_i|\mathbf{x}) \end{aligned}$$

As might be intuitively expected we see that the optimal decision in this case is to choose the model with the highest posterior probability.

It can be shown that, under the "computer game scenario" mentioned earlier, $P(M_i|\mathbf{x}) \to 1$ for the "true" model as $n \to \infty$, meaning that the Bayes procedure is *consistent*.

In the case where only two models (say M_1 and M_2) are available for comparison, a measure of plausibility is the *Bayes factor* specified below.

Definition 6 Bayes factor. *Given two models M_1 and M_2 for data $\mathbf{x}$, the Bayes factor in favor of M_1 (and against M_2) is give as the posterior to prior odds ratio*

$$B_{12} = \frac{p(\mathbf{x}|M_1)}{p(\mathbf{x}|M_2)} = \frac{P(M_1|\mathbf{x})}{P(M_2|\mathbf{x})} / \frac{P(M_1)}{P(M_2)} \tag{89}$$

Intuitively, the Bayes factor says whether the data have increased ($B_{12} > 1$) or decreased ($B_{12} < 1$) the odds on M_1. Clearly, if the prior weights are uniform, the Bayes factor is simply a ratio of the posterior weights. A thorough discussion about the properties and guidelines for interpretation of the Bayes factor can be found in Kass and Raftery (1995).

Example with Bayes factor

Consider again the thumbtack tossing or the cigar box sampling problems discussed earlier.

We proceed now by assuming that two sequences of n_1 and n_2 binary observations are made, respectively, and there are two potential models for the data.

Each of the sequences is such that it could potentially be modeled with the thumbtack tossing scenario, however, the thumbtack properties might be different in the two situations.

M_1 states that all $n_1 + n_2$ observations are exchangeable.

M_2 states that the two sequences of n_1 and n_2 observations are separately exchangeable, but not when combined.

This can be given an operational interpretation that the generating probability distribution is distinct for the two sequences under M_2 and the same under $M_{1.}$

The operational parameters for M_2 are labeled as θ and ψ, whereas for M_1 we only have a single parameter, say θ.

Given M_1, the predictive probability of the data can be written as

$$p(x_1, ..., x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2}|M_1) = \int_0^1 \prod_{i=1}^{n_1+n_2} \theta^{x_i}(1-\theta)^{1-x_i} p(\theta) d\theta. \tag{90}$$

If the prior is defined equal to the conjugate Beta(α, β) distribution as previously, then the predictive distribution will take the explicit form

$$\begin{aligned} & p(x_1, ..., x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2}|M_1) \\ = & \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=1}^{n_1+n_2} x_i)\Gamma(\beta+n_1+n_2-\sum_{i=1}^{n_1+n_2} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_1+n_2)}. \end{aligned}$$

Under model M_2 it is necessary to specify two distinct prior distributions, one for both parameters θ and ψ.

The predictive distribution of the data is then

$$\begin{aligned} & p(x_1, ..., x_{n_1}|M_2)p(x_{n_1+1}, ..., x_{n_1+n_2}|M_2) \\ = & \int_0^1 \prod_{i=1}^{n_1} \theta^{x_i}(1-\theta)^{1-x_i} p(\theta)d\theta \times \\ & \times \int_0^1 \prod_{i=n_1+1}^{n_1+n_2} \psi^{x_i}(1-\psi)^{1-x_i} p(\psi)d\psi. \end{aligned} \tag{91}$$

If the Beta(α, β) prior is assigned to both θ and ψ, the predictive distribution can be written as

$$\begin{aligned} & p(x_1, ..., x_{n_1}|M_2)p(x_{n_1+1}, ..., x_{n_1+n_2}|M_2) \\ = \ & \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=1}^{n_1} x_i)\Gamma(\beta+n_1-\sum_{i=1}^{n_1} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_1)} \times \\ & \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=n_1+1}^{n_1+n_2} x_i)\Gamma(\beta+n_2-\sum_{i=n_1+1}^{n_1+n_2} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_2)}. \end{aligned}$$

Finally, the Bayes factor for model comparison now equals

$$\frac{p(x_1, ..., x_{n_1}, x_{n_1+1}, ..., x_{n_1+n_2}|M_1)}{p(x_1, ..., x_{n_1}|M_2)p(x_{n_1+1}, ..., x_{n_1+n_2}|M_2)}$$

$$= \frac{\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=1}^{n_1+n_2} x_i)\Gamma(\beta+n_1+n_2-\sum_{i=1}^{n_1+n_2} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_1+n_2)}}{\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=1}^{n_1} x_i)\Gamma(\beta+n_1-\sum_{i=1}^{n_1} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_1)} \times \\ \times \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+\sum_{i=n_1+1}^{n_1+n_2} x_i)\Gamma(\beta+n_2-\sum_{i=n_1+1}^{n_1+n_2} x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_2)}}.$$

Behavior of this Bayes factor will be examined during the lectures using 'the cigar box simulation scenario'.

Also, one course exercise focuses on investigating the behavior as a function of the prior hyperparameters.

The previously mentioned approach focusing on choosing a single model as representation for data as such is not the most sensible solution in all situations, especially if we are aiming to produce some kind of statements about observables using our models (*e.g.* prediction of future values).

An optimal strategy under such circumstances does not even necessitate a choice of a model, which can be formalized using a different utility structure.

Let a be an answer relating to the unknown of interest ω.

The answer can for instance be a value of a future observation or an estimate of parameter common to all models in $\mathcal{I}$.

With the utility function $u(a, \omega)$, the expected utility of an answer a^* becomes

$$\bar{u}(a^*|\mathbf{x}) = \int u(a^*, \omega) p(\omega|\mathbf{x}) d\omega \tag{92}$$

and this is indeed the optimal answer if

$$\bar{u}(a^*|\mathbf{x}) = \sup_a \bar{u}(a|\mathbf{x}) \tag{93}$$

Note that $p(\omega|\mathbf{x})$ still has the posterior weighted mixture form.

This type of a strategy is often called Bayesian model averaging, and its sensibility for the problem at hand is dependent on whether the chosen utility function reflects the relevant issues (which ultimately need to be considered by the modeler).

An important question yet to be stated is that: What happens if all models in $\mathcal{I}$ are poor descriptions of $\mathbf{x}$**?**

It should be clear from the above that we cannot use the previously stated formalism to directly detect this.

Notice that the crucial question in such a situation is not the comparison of models, but the criticism of a model *without suggesting anything to replace it*.

If one had, for instance, a more general model in mind than those included in $\mathcal{I}$, the problem would be easily resolved by taking that model also into consideration and proceeding as before.

Indeed, in many situations a generalization of the finite-dimensional parametric models $p(\mathbf{x}|\boldsymbol{\theta})$ could be obtained by considering models involving directly a probability measure on the space of distribution functions (these are typically called non-parametric models in the Bayesian framework).

Including such models to the class I and then performing the formal analysis, can be seen as one promising strategy to check formally the plausibility of the various parametric assumptions (*e.g.* Gutiérrez-Peña and Walker, 2001).

Example 22 Genetic differentiation among populations. *We consider investigation of the genetic similarity of some geographically separated populations using the approach of Corander et al. (2003, 2004). We invoke a sampling design where individuals are gathered from N_P distinct populations based on available prior knowledge concerning their geographical separation. Assume that genotypes are observed at N_L independent marker loci (meaning e.g. that they are located in different chromosomes), where at each locus j there are $N_{A(j)}$ possible alleles to be distinguished.*

Since the true underlying population substructure is unknown, the number of populations with differing allele frequencies is treated as a parameter ν, having the range of reasonable values $[1, N_P]$ where the upper bound is directly given by the sampling design.

In the sequel, a population refers thus to a genetic source having allele frequencies distinct from other sources.

At locus j, the unobserved probability of observing allele A_{jk} (allele frequency) in population i is represented by p_{ijk} $(i = 1, ..., \nu;\ j = 1, ..., N_L;\ k = 1, ..., N_{A(j)})$.

To simplify the notation, θ will be used as a generic symbol jointly for the allele frequencies (θ_i for population i).

Similarly n will represent jointly the observed marker allele counts n_{ijk}.

The partition of the original populations can be represented by a parameter

$S = (s_1, ..., s_\nu)$, where $s_i, i = 1, ..., \nu$, contains the indices of the sample populations deemed to have equal allele frequencies.

The joint distribution of the observed marker allele counts and the model parameters is specified by

$$\begin{aligned} \pi(\theta,\nu, S, n) &= \pi(n|\theta, \nu, S)\pi(\theta|\nu, S)\pi(S|\nu)\pi(\nu) \qquad (94) \\ &\propto \prod_{i=1}^{\nu} \prod_{j=1}^{N_L} \prod_{k=1}^{A(j)} \left[p_{ijk}^{n_{ijk}} \pi(p_{ijk})\right] \pi(S|\nu)\pi(\nu), \end{aligned}$$

where

$$\pi(n|\theta,\nu, S) \propto \prod_{i=1}^{\nu} \prod_{j=1}^{N_L} \prod_{k=1}^{A(j)} p_{ijk}^{n_{ijk}} \qquad (95)$$

is the multinomial likelihood,

$$\pi(\theta|\nu, S) = \prod_{i=1}^{\nu} \prod_{j=1}^{N_L} \prod_{k=1}^{A(j)} \pi(p_{ijk}) \tag{96}$$

is the prior density of θ, and

$$\pi(S|\nu)\pi(\nu) \tag{97}$$

is the joint prior of the structure parameters.

Notice that when the allele frequencies of two original populations are stated to be equal in the model (their indices belong to the same s_i), their observed counts in n can be summed together in the likelihood.

If the prior beliefs about allele frequencies are represented by the Dirichlet distribution with hyperparameter λ_{ijk}, then, for a fixed value of (ν, S), the joint distribution of the data under our predictive probability model equals

$$\pi(n|\nu_P, S) = \int \pi(n|\theta)\pi(\theta)d\theta \tag{98}$$

$$= \prod_{i=1}^{\nu_P} \prod_{j=1}^{N_L} \frac{\Gamma(\sum \lambda_{ijk})}{\Gamma(\sum \lambda_{ijk} + n_{ijk})} \prod_{k=1}^{A(j)} \frac{\Gamma(\lambda_{ijk} + n_{ijk})}{\Gamma(\lambda_{ijk})} \tag{99}$$

The above model arises theoretically from the assumption of specific generalized exchangeability (Corander *et al.* 2007, 2009).

In particular, they combined this property with the reference prior having $\lambda_{ijk} = 1/N_{A(j)}, k = 1, ..., N_{A(j)}$, which dates back already to Perks (1947).

Example 23 Genetic differentiation among populations (continued). *In the earlier considered example concerned with genetic differentiation, we obtained a predictive model for the observed allele counts given the structure parameter S (specifying the groups of populations with different allele frequencies). Let the predictive model be abbreviated as $\pi(n|S)$. The posterior distribution of the structure parameter thus becomes*

$$\pi(S|n) = \frac{\pi(n|S)\pi(S)}{\sum_{S \in \mathcal{S}} \pi(n|S)\pi(S)} \tag{100}$$

For small values of N_P we may calculate the posterior probabilities exactly by exhaustive enumeration.

The number of distinct values of S (i.e. partitions of the finite set $\{1, ..., N_P\}$) equals the sum $\sum_{\nu_P=1}^{N_P} \sigma_{\nu_P}^{N_P}$ where $\sigma_{\nu_P}^{N_P}$ is the Stirling number of the second kind.

For example, for $N_P = 10$, we get $\sum_{\nu_P=1}^{N_P} \sigma_{\nu_P}^{N_P} = 115{,}975$.

For moderate or large values of N_P, simulation techniques need to be used for estimation of the posterior probabilities.

In this genetics example it is sometimes also natural to consider model averaging over the posterior distribution S. When one is interested in a quantity depending on the allele frequencies θ, such as the degree of genetic differentiation among the populations, its posterior distribution is obtained by averaging the conditional posterior distributions of p_{ijk} over (100).

Model comparison, Part III

In this section we investigate some concepts of frequentist statistical analysis in the context of model comparison.

In particular, we shall see some connections with the Bayesian approach introduced in the previous section.

The classical Neyman-Pearson theory for testing models requires pairwise processing of the elements of a model class I, and therefore, let us concentrate for a moment on the situation where $\mathcal{I}$ contains only two models: M_1 and M_2.

A generally accepted device for comparing models' appropriateness for a particular data set $\mathbf{x}$ is the likelihood ratio

$$\frac{p(\mathbf{x}|M_1)}{p(\mathbf{x}|M_2)}, \tag{101}$$

which is identical to the Bayes factor in the case of *completely specified models* (no parameters are estimated) and equal prior probabilities.

Typically, however, models contain unknown parameters and the frequentist comparison procedure differs from the Bayes factor.

Using Neyman-Pearson theory we formulate the null hypothesis H_1 : the observations have arisen from the model M_1, and the alternative H_2 : the observations have arisen from the model M_2.

To be able to formulate a regular likelihood ratio test (see *e.g.* Cox and Hinkley, 1974) of H_1 against H_2, assume the ”nested hypothesis” case where M_2 is the full model and M_1 a reduced version of M_2 where some parameter(s) have been given fixed values.

Let $d(\boldsymbol{\theta}_i)$ generally denote the number of unrestricted parameters in $M_i, i \in \mathcal{I}$.

The likelihood ratio test is formulated as: reject H_1 if

$$\lambda_n = \frac{L(\hat{\boldsymbol{\theta}}_1|\mathbf{x})}{L(\hat{\boldsymbol{\theta}}_2|\mathbf{x})} < c < 1 \tag{102}$$

where c is *a priori* specified threshold and $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i, i \in I$.

We notice the difference with the Bayes factor where the uncertainty about parameters is accounted for by integrating them out with respect to the prior distribution, instead of maximization.

Under general regularity conditions on $L(\boldsymbol{\theta}_i|\mathbf{x})$ (*e.g.* $d(\boldsymbol{\theta}_i)$ remains fixed as $n \to \infty$), $-2\log\lambda_n$ is approximately chi-square distributed with $d(\boldsymbol{\theta}_2)-d(\boldsymbol{\theta}_1)$ degrees of freedom (denoted by $\chi^2_{d(\boldsymbol{\theta}_2)-d(\boldsymbol{\theta}_1)}$).

As illustrated in Gelfand and Dey (1994), an inconsistency of this procedure is evident, since

$$\begin{aligned}\lim_{n\to\infty}\{P(\text{choose } M_2|M_1 \text{ true})\} &= \lim_{n\to\infty}\{P(\lambda_n < c|M_1 \text{ true})\} \qquad (103)\\ &= \lim_{n\to\infty}\{P(-2\log\lambda_n > -2\log c)\}\\ &= P(\chi^2_{d(\boldsymbol{\theta}_2)-d(\boldsymbol{\theta}_1)} > -2\log c) > 0\end{aligned}$$

Thereby, even with unlimited amounts of data the procedure is not guaranteed to pick out the correct model.

A more severe problem associated with the above testing scenario is that it provides no general yardstick for comparison of a range of different models.

For instance, when the evidence against each of the models in $\mathcal{I}$ is measured by the *p*-value according to (102) where the unrestricted model M_2 is the most general model in $\mathcal{I}$, it follows that the *p*-value is a decreasing function of the number of restrictions imposed on $\boldsymbol{\theta}$.

Thereby, the largest possible model is by definition associated with a *p*-value equal to unity, while the remaining models attain *p*-values smaller than or equal to unity depending on their degree of fit to data with respect to the full model.

Generally, this framework makes especially the comparison of non-nested models difficult.

Hypothesis tests are designed to detect *any* discrepancies between a model and reality.

Since models are virtually never exact descriptions of reality, we know by definition that for large enough samples the discrepancies will be detected by (102) and lead to a rejection of M_1 even if it is a good model for the purpose at hand.

The point is that rejection of M_1 does not necessarily mean that M_2 offers a better description of the data, and hence, one should *compare* the two models instead of simply looking at the discrepancy between M_1 and the data.

In this respect, a fundamental flaw of the hypothesis test scenario is that it cannot provide directly evidence *for* a model but only *against* it.

Even some of the advocates of the frequentist approach to statistical inference have clearly pointed out that such framework is unfortunate in the context of model selection and suggested that other approaches, such as those discussed in the following section, should preferably be followed (*e.g.* see Lindsey, 1996).

Model comparison, Part IV

Recall from the section where reference priors were introduced, that the prior opinion for a mathematically derived formula is often expressed as

$$p(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta}) \tag{104}$$

where the proportionality meant that the prior distribution is understood as proportional to a real-valued function $g(\boldsymbol{\theta})$ for which $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})d\boldsymbol{\theta} \neq 1$, but which satisfies

$$\int p(\mathbf{x}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta} < \infty \tag{105}$$

Under this assumption the "posterior"

$$\frac{p(\mathbf{x}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{106}$$

is a well-defined probability measure, since the unknown proportionality constant cancels in the above formula.

In the model comparison framework the improper priors introduce a more severe problem, since the above "logic of proportionality" is not applicable to the situation where $\boldsymbol{\theta}_i$ have varying dimensions.

This is due to the fact that the imagined constants do not cancel, *e.g.* in (89).

To resolve this problem, a wide range of approaches has been suggested in the statistical literature, see *e.g.* Key *et al*. (1999).

Here we review some of the more prominent proposals.

For simplicity, consider again the case where I contains only two models M_1 and M_2.

Suppose that the data $x_1, ..., x_n$ is split into two parts such that

$$x_1, ..., x_m, x_{m+1}, ..., x_n = (\mathbf{y}, \mathbf{z}) = \mathbf{x} \tag{107}$$

Let $\mathbf{y}$ be a *training data* set.

For an improper prior of the above type, the *partial* Bayes factor (89) for M_1 against M_2 based on $\mathbf{z}$ after the training data $\mathbf{y}$, equals

$$\begin{aligned} B_{12}(\mathbf{z}|\mathbf{y}) &= \frac{p(\mathbf{z}|M_1, \mathbf{y})}{p(\mathbf{z}|M_2, \mathbf{y})} \\ &= \frac{\int p(\mathbf{z}|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1|\mathbf{y})d\boldsymbol{\theta}_1}{\int p(\mathbf{z}|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2|\mathbf{y})d\boldsymbol{\theta}_2} \end{aligned} \tag{108}$$

which is well-defined if the posteriors

$$p(\boldsymbol{\theta}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)}{\int p(\mathbf{y}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}, i = 1, 2 \tag{109}$$

are proper densities.

While this approach seems to resolve the indeterminacy problem, it is unsatisfactory in the sense that it is dependent on the arbitrary division of the data.

Also, the question of how much data should be used in the training has to be considered.

We may consider a training set $\mathbf{y}$ as a *proper* one if the integrals in (108) converge.

Then, if no subset of $\mathbf{y}$ is proper, the training set may be called *minimal*.

While this concept is useful in some problems, it should be noted, however, that for most discrete data problems minimal training sets are not defined, which greatly limits the applications.

The *intrinsic Bayes factors* of Berger and Pericchi (1996) are based on averaging the partial Bayes factors over all possible minimal training sets.

O'Hagan (1995), in turn, introduced an idea which is based on using a minimal *fraction* $0 < b < 1$ of the likelihood as a training set, which leads to

$$p(\mathbf{x}|M_i, b) = \frac{\int p(\mathbf{x}|\boldsymbol{\theta}_i)g(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}{\int \{p(\mathbf{x}|\boldsymbol{\theta}_i)\}^b g(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i} \tag{110}$$

Since $b = m/n$, the likelihood $p(\mathbf{y}|\boldsymbol{\theta}_i)$ is approximately the full likelihood $p(\mathbf{x}|\boldsymbol{\theta}_i)$ raised to the power b (notice that this requires the likelihood to have the product form with n conditionally independent terms).

This procedure corresponds to using a fraction of the average information in the data, and also leads to a well-defined "partial" Bayes factor.

Apart from the two approaches mentioned above, there is a wealth of more or less related "automated" methods for Bayesian model comparison.

As was seen in the previous section, the frequentist approach to model selection via hypothesis testing typically uses the large sample behavior of the likelihood ratio.

Similar ideas may be fruitfully pursued in the Bayesian framework.

Here we intend to give a heuristic derivation of the central results without focusing on the somewhat involved technical details.

Model comparison, Part V

Consider the parametric case with a model labeled by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for an exchangeable sequence of observations. We then have

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \\ &\propto \exp\{\log p(\boldsymbol{\theta}) + \log p(\mathbf{x}|\boldsymbol{\theta})\} \end{aligned} \tag{111}$$

Let $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_n$ denote the respective maxima of the two logarithmic terms in (111), *i.e.* the prior mode and the maximum likelihood estimate, respectively.

These are determined by setting $\nabla \log p(\boldsymbol{\theta}) = 0$ and $\nabla \log p(\mathbf{x}|\boldsymbol{\theta}) = 0$, respectively.

By expanding both logarithmic terms about their respective maxima we obtain

$$\begin{aligned}\log p(\boldsymbol{\theta}) &= \log p(\hat{\boldsymbol{\theta}}_0) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0)' H(\hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0) + R_0 \qquad (112)\\ \log p(\mathbf{x}|\boldsymbol{\theta}) &= \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + R_n\end{aligned}$$

where R_0, R_n denote remainder terms and

$$\begin{aligned}H(\hat{\boldsymbol{\theta}}_0) &= \left(-\frac{\partial^2 \log p(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0} \qquad (113)\\ H(\hat{\boldsymbol{\theta}}_n) &= \left(-\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}\end{aligned}$$

are the Hessian matrices.

Under regularity conditions which ensure that the remainder terms R_0, R_n are small for large n, we get the result

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)'H(\hat{\boldsymbol{\theta}}_0)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'H(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\right\} \tag{114}$$

The Hessian matrix $H(\hat{\boldsymbol{\theta}}_n)$ measures the local curvature of the log-likelihood function at it maximum $\hat{\boldsymbol{\theta}}_n$ and is typically called the *observed information matrix*.

Further, by ignoring the prior terms (which are swamped by the data as n grows) we see that the posterior can be approximated by the multivariate normal distribution with mean $\hat{\boldsymbol{\theta}}_n$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_n = H(\hat{\boldsymbol{\theta}}_n)^{-1}$.

However, asymptotics also reveal that

$$\begin{aligned}\lim_{n\to\infty}\left\{\frac{1}{n}\left(-\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} &= \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{l=1}^{n}\left(-\frac{\partial^2 \log p(x_l|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)\right\} \qquad (115)\\ &= \int p(x|\boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx\end{aligned}$$

so that $H(\hat{\boldsymbol{\theta}}_n) \to n\mathbf{I}(\hat{\boldsymbol{\theta}}_n)$, where $\mathbf{I}(\boldsymbol{\theta})$ is (again) the *Fisher information matrix*, defined as

$$(\mathbf{I}(\boldsymbol{\theta}))_{ij} = \int p(x|\boldsymbol{\theta})\left(-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right)dx \qquad (116)$$

The above results can be utilized in the model comparison framework through an approximation to the key quantity $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$, the marginal likelihood.

An important assumption concerning the validity of the asymptotic approximation is that the dimension $d(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ remains fixed as $n \to \infty$.

Using the properties of the multivariate normal distribution (*i.e.* the form of its normalizing constant), an approximation to the marginal likelihood can be written as

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (117) \\ &\approx (2\pi)^{d(\hat{\boldsymbol{\theta}}_n)}|\hat{\boldsymbol{\Sigma}}_n|^{1/2}p(\hat{\boldsymbol{\theta}}_0)p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) \end{aligned}$$

Using this approximation the posterior weights of the different models in I can be calculated.

Under the assumption that the prior is continuous in $\boldsymbol{\Theta}$ and bounded at $\hat{\boldsymbol{\theta}}_0$, an approximate Bayes solution to the model comparison problem under the 0-1 loss scheme given in (85), is to choose the model which maximizes

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) + \frac{1}{2} \log |\hat{\boldsymbol{\Sigma}}_n| + d(\hat{\boldsymbol{\theta}}_n) \log(2\pi) \tag{118}$$

This result is valid under a rather general setting (see Kim, 1998) and defines a consistent model selection procedure.

However, a yet simpler and *still consistent* model comparison criterion is obtained, when terms not depending on n are ignored, and an asymptotic expansion of $\log |\hat{\boldsymbol{\Sigma}}_n|$ is used.

Under certain conditions (see Kim, 1998) the log-determinant can be written as

$$\log |\hat{\boldsymbol{\Sigma}}_n| = -2 \log \left(\prod_{l=1}^{d(\hat{\boldsymbol{\theta}}_n)} s_l(n) \right) + R_0 \tag{119}$$

where the remainder is bounded in n and the terms $s_l(n)$ are the *rates of convergence* of the maximum likelihood estimate $\hat{\theta}_{l(n)}$ to the true value of the (l)th component θ_l of $\boldsymbol{\theta}$.

Under regular $\sqrt{n}$-convergence we are led to the criterion

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \log\left(\prod_{l=1}^{d(\hat{\boldsymbol{\theta}}_n)} n^{1/2}\right) = \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \frac{d(\hat{\boldsymbol{\theta}}_n)}{2}\log n \qquad (120)$$

This is precisely the widely-known criterion derived by Schwarz (1978), often called BIC or SBC (sometimes the above is multiplied by two).

In the two model case, we can more concretely write

$$\log B_{12} \approx \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{1(n)}) - \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{2(n)}) - \frac{d(\hat{\boldsymbol{\theta}}_{1(n)}) - d(\hat{\boldsymbol{\theta}}_{2(n)})}{2}\log n \qquad (121)$$

Although (120) is a rather rough approximation, it can generally be considered as guideline for model comparison in a situation where the prior information is vague and difficult to specify precisely.

Notice that also from (120) one can derive approximate posterior weights for the elements of $\mathcal{I}$.

Generally, the criterion (120) has in various simulation studies shown to be conservative, such that for small n it may underestimate the true model dimension.

Since the introduction of the model comparison criterion AIC by Akaike (1974), a considerable interest has been attained in the statistical literature to criteria of the *penalized maximum likelihood* type

$$\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - c \cdot d(\hat{\boldsymbol{\theta}}_n) \log g(n) \tag{122}$$

where different choices of c and $g(n)$ lead to different suggested criteria.

For instance, $c = 1$ and $g(n) = e^1$ give rise to the AIC, $c = 1$ and $g(n) = \log n$ to the criterion of Hannan and Quinn (1979), and $c = 1/2$ and $g(n) = n$ to (120).

It can be shown that for problems where $d(\hat{\boldsymbol{\theta}}_n)$ is *not increasing* with n, any choice of $g(n)$ equal to a constant, will lead to an inconsistent criterion.

In particular, AIC is not consistent, and it typically leads to a gross overestimation of a reasonable dimension of $\boldsymbol{\theta}$ when n is large.

On the other hand, AIC and criteria alike it can have better future predictive performance in situations where the model structure itself is not of interest.

Computational tools for Bayesian modelling

Computer simulation forms a central part of Bayesian data analysis, since it greatly enhances our possibilities to fit useful (often complicated) models to data.

A set of realizations from a posterior distribution also provides a convenient means to specify numerical inference summaries, such as histograms, means, quantiles etc.

One prominent computational strategy for Bayesian modelling is known as Markov chain Monte Carlo (MCMC), see *e.g.* Gilks *et al.* (1996), Robert and Casella (1999, 2004).

Some of the most popular MCMC algorithms are known as the Gibbs sampler, Metropolis-Hastings, and reversible jump Metropolis-Hastings algorithms.

There exists a huge MCMC literature, and there are hundreds of different variants of these and other algorithms available for Bayesian computation.

Let the Bayesian joint model for the data and parameters be specified as:

$$p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{123}$$

From the joint model, we get the posterior as

$$\frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{124}$$

Further, let the parameters be divided into m groups, such that

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_m). \tag{125}$$

In general, any component $\theta_i \in \boldsymbol{\Theta}_i$ of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ may be considered as a scalar, vector, or matrix.

The purpose of the (single component) Metropolis-Hastings (MH) algorithm is to simulate a Markov chain, which has the stationary distribution equal to

(124).

This may be done as follows.

Let $\boldsymbol{\theta}^0$ be an initial value of the parameter $\boldsymbol{\theta}$.

Further, let $\boldsymbol{\theta}^t$ denote the tth value of $\boldsymbol{\theta}$ in the Markov chain $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots$, with transition between different states (here from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^*$) governed by the acceptance ratio

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*) = \min\left(1, \frac{p(\mathbf{x}|\boldsymbol{\theta}^t_{-i}, \theta^*_i)p(\boldsymbol{\theta}^t_{-i}, \theta^*_i)}{p(\mathbf{x}|\boldsymbol{\theta}^t_{-i}, \theta^t_i)p(\boldsymbol{\theta}^t_{-i}, \theta^t_i)} \frac{q_{-i}(\boldsymbol{\theta}^t_i|\boldsymbol{\theta}^t_{-i}, \theta^*_i)}{q_{-i}(\boldsymbol{\theta}^*_i|\boldsymbol{\theta}^t_{-i}, \theta^t_i)}\right), i = 1, \ldots, m, \tag{126}$$

where

$$q_{-i}(\cdot|\boldsymbol{\theta}^t_{-i}, \theta^t_i) \tag{127}$$

is a so called proposal distribution for the component θ_i, given the current values of the remaining components of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^t_{-i}$, and also θ^t_i.

An example of a proposal distribution is a multivariate normal distribution with mean θ^t_i and fixed covariance matrix $\boldsymbol{\Sigma}$.

In practice, the MH algorithm typically proceeds as:

At a given t (often called a sweep), for $i = 1, ..., m$:

1. Simulate $\theta_i^* \sim q_{-i}(\cdot|\boldsymbol{\theta}^t_{-i}, \theta_i^t)$
2. Simulate $u \sim Uniform(0, 1)$
3. If $u \leq \alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*)$, set $\theta_i^t = \theta_i^*$, else keep θ_i^t.

A special case of the single component Metropolis-Hastings algorithm is the Gibbs sampler.

In Gibbs sampler, the proposal distribution is the *full conditional* distribution, which is not dependent of the current value θ_i^t,

$$q_{-i}(\cdot|\boldsymbol{\theta}_{-i}^t, \theta_i^t) = \frac{p(\mathbf{x}|\boldsymbol{\theta}_{-i}^t, \theta_i)p(\boldsymbol{\theta}_{-i}^t, \theta_i)}{\int p(\mathbf{x}|\boldsymbol{\theta}_{-i}^t, \theta_i)p(\boldsymbol{\theta}_{-i}^t, \theta_i)d\theta_i}. \tag{128}$$

Notice that the normalizing constant cancels in the acceptance ratio.

For the Gibbs sampler each proposal value will thus be accpted as the acceptance ratio equals always 1.

Various sampling techniques may be utilized for generating values from the full conditional distributions (direct sampling, rejection sampling etc).

In many applications the dimension of $\boldsymbol{\theta}$ varies between different putative models we are interested in, so that $\boldsymbol{\theta}^*$ may have a larger or smaller dimension than $\boldsymbol{\theta}^t$.

In the so called reversible jump (RJ) MCMC algorithm introduced by Green (1995), the dimension switch is handled by a transformation of variables such that $\boldsymbol{\theta}^*$ is a deterministic function of $\boldsymbol{\theta}^t$ and a realization of another random variable, generated from a distribution specified in the setup of the proposal mechanism.

Often the transformation is such that the proposal densities involve a Jacobian that needs to be calculated explicitly.

The RJ MCMC approach has been widely adopted for the calculation of posterior probabilities of models.

From the realization of such MCMC simulation it is possible to obtain an approximation to the posterior probabilities (84) by the relative number of visits in the chain to a specific model M_i.

Let $M^t \in \mathcal{I}$ be the index of the model at t, then

$$\hat{P}(M_i|\mathbf{x}) = n^{-1} \sum_{t=1}^{n} I(M^t = M_i) \qquad (129)$$

is a consistent estimate of the posterior probability of M_i.

Also, the model averaging idea is crystallized in the variable dimensional MCMC, where, *e.g.* predictions can be generated at each t.

The model structures which have most support from the data, will then be most often used to generate the predictions.

Example 24 Genetic differentiation among populations (continued). *In the earlier considered example concerned with genetic differentiation, we obtained a predictive model for the observed allele counts given the structure parameter S (specifying the groups of populations with different allele frequencies). The posterior distribution of the structure parameter was specified as*

$$\pi(S|n) = \frac{\pi(n|S)\pi(S)}{\sum_{S\in\mathcal{S}} \pi(n|S)\pi(S)} \tag{130}$$

Since we are in general unable to use complete enumeration to calculate

$$\sum_{S\in\mathcal{S}} \pi(n|S)\pi(S), \tag{131}$$

MCMC provides a useful strategy for the estimation problem.

Corander *et al.* (2004a) introduced a Metropolis-Hastings algorithm, which simulates a Markov chain having the stationary distribution equal to the above posterior.

In fact, they investigated a slightly more general problem, where the sample populations were also allowed to consist of a single individual only.

Their MH algorithm is defined by the transition kernel, which determines the probability of a transition from a current state S to a proposed new state S^*, as

$$\min\left(1, \frac{\pi(n|S^*)}{\pi(n|S)}\frac{q(S|S^*)}{q(S^*|S)}\right), \tag{132}$$

where the $\pi(n|S)$ (marginal likelihood) is defined according to

$$\prod_{i=1}^{\nu_P}\prod_{j=1}^{N_L}\frac{\Gamma(\sum\lambda_{ijk})}{\Gamma(\sum\lambda_{ijk}+n_{ijk})}\prod_{k=1}^{A(j)}\frac{\Gamma(\lambda_{ijk}+n_{ijk})}{\Gamma(\lambda_{ijk})} \tag{133}$$

$q(S^*|S)$ is the probability of choosing state S^* as the candidate for the next state when in S

$q(S|S^*)$ is the probability of restoration of the current state S.

The proposal mechanism to derive S^* from S considered by Corander *et al.* (2004) was constructed from the following four different possibilities:

- With probability 1/2, merge two randomly chosen classes s_c, s_{c*}.
- With probability 1/2 split a randomly chosen class s_c into two new classes, whose cardinalities are uniformly distributed between 1 and $|s_c| - \mathbf{1}$, and whose elements are randomly chosen from s_c.
- Move an arbitrary item from a randomly chosen class $s_c, |s_c| > \mathbf{1}$, into another randomly chosen class s_{c^*}.
- Choose one item randomly from each of two randomly chosen classes s_c and s_{c^*}, and exchange them between the classes.

Assuming that S consists of k classes, the respective proposal probabilities $q(S^*|S)$ corresponding to these different transition types can be written as:

$$\begin{aligned}
1 &: \binom{k}{2}^{-1}/2 \\
2 &: \begin{cases} k^{-1}\lfloor|s_c|/2\rfloor^{-1}\binom{|s_c|}{|s_{c^*}|}^{-1}/2, \text{ if } |s_{c^*}| < |s_c|/2 \\ k^{-1}\lfloor|s_c|/2\rfloor^{-1}\binom{|s_c|}{|s_{c^*}|}^{-1}/4, \text{ if } |s_{c^*}| = |s_c|/2 \end{cases} \\
3 &: \tau(S)^{-1}(k-1)^{-1}|s_c|^{-1} \\
4 &: \binom{k}{2}^{-1}|s_c|^{-1}|s_{c^*}|^{-1},
\end{aligned} \tag{134}$$

where $s_{\acute{c}^*}$ in the second transition type is one of the two new classes formed by splitting s_c, having the smallest cardinality $|s_{c^*}|$.

Further, $\tau(S)$ in the third transition type is the number of classes with $|s_c| > 1$.

When the current state S is such that not all move types are available, trivial changes will be imposed on the proposal probabilities.

The above specified transition mechanism defines an aperiodic and irreducible finite Markov chain.

Since the state space $\mathcal{S}$ of the chain is finite, it follows (*e.g.* Häggström, 2002) that (132) defines also a positive recurrent Markov chain.

Thus, for a realization of the chain $\{S_t, t = 0, 1, ...\}$ we have

$$\lim_{n \to \infty} p_n(S|n) = p(S|n), \tag{135}$$

where

$$p_n(S|n) = n^{-1} \sum_{t=0}^{n} I(S_t = S) \tag{136}$$

is the relative frequency of occurrence of state S.

Clearly, the optimal Bayesian classification is obtained from the realization $\{S_t, t = 0, 1, ...\}$ as the value of S, which maximizes the marginal likelihood.

However, convergence of the chain may be slow in reality for large spaces $\mathcal{S}$, and therefore, Corander *et al*. (2004a) proposed an alternative estimate

$$p_n^*(S|n) = \frac{\pi(n|S)}{\sum_{S \in \mathcal{S}^*} \pi(n|S)}, \tag{137}$$

where $\mathcal{S}^*$ is the set of distinct values of S visited in m independent realizations of the Markov chain $\{S_{tj}, t = 0, 1, ...; j = 1, ..., m\}$.

Under the stated conditions, (137) is also a consistent estimate of $p(S|n)$.

The advantages of using this estimate rather than (136) are the relative stability of the sum $\sum_{S \in \mathcal{S}^*} \pi(n|S)$, and that results from independent realizations of the chain can be joined in a meaningful way, which is not possible for (136).

In the relative frequency based estimation, some chains may become stuck in regions of $\mathcal{S}$ associated with low values of $\pi(n|S)$, and obtain thereby too much weight in the estimate (136).

In typical applications the number of chains m needed for reliable estimates ranges from some dozens to several hundreds. Corander *et al*. (2004a) used a parallel solution to produce the chains, in order to aid in visual inspection of the convergence of the estimation procedure.

References

[1] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Auto. Control,* **19**, 716-723.

[2] Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Stat. Assoc.* **91**, 109-122.

[3] Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian theory. Chichester: Wiley.

[4] Corander, J., Waldmann, P. and Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367-374.

[5] Corander, J., Waldmann, P., Marttinen, P. and Sillanpää, M. J. (2004a). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363-2369.

[6] Corander, J., Gyllenberg, M. and Koski, T. (2007). Random partition models and exchangeability for Bayesian identification of population structure. *Bull. Math. Biol.* **69**, 797-815.

[7] Corander, J., Gyllenberg, M. and Koski, T. (2009). Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Advances in Data Analysis and Classification, **3**, 3-24.

[8] Cox, D. R. and Hinkley, D. V. (1974). Theoretical statistics. London: Chapman&Hall.

[9] de Finetti, B. (1974). Theory of probability **1**. Chichester: Wiley.

[10] Felsenstein, J. (2004). Inferring phylogenies. Sunderland: Sinauer Associates.

[11] Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc.*, **B 56**, 501-514.

[12] Gutiérrez-Peña, E. and Walker, S. G. (2001). A Bayesian predictive approach to model selection. *J. Statist. Plannig Inference*, **93**, 259-276.

[13] Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc.*, **B 41**, 190-195.

[14] Kass, R. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Stat. Assoc.* **90**, 773-795.

[15] Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Stat. Assoc.* **91**, 1343-1370.

[16] Key, J. T., Pericchi, L. R. and Smith, A. F. M. (1999). Bayesian model choice: What and why? In Bernardo, J., Berger, J., Dawid, A. and Smith,

A. (Eds.). Bayesian Statistics 6, Oxford: Oxford University Press, 343-370.

[17] Kim, J-Y. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica*, **66**, 359-380.

[18] Lindsey, J. (1996). Parametric statistical inference. Oxford: Oxford University Press.

[19] Murray, R.G., McKillop, J. H., Bessant, R. G. Hutton, I., Lorimer, A. R. and Lawrie, T. D. V. (1981). Bayesian analysis of stress thallium-201 scintigraphy. *Eur. J. Nucl. Med.,* **6**, 201-204.

[20] O'Hagan, A. (1994). Kendall's advanced theory of statistics. Vol. 2B: Bayesian inference. London: Edward Arnold.

[21] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J.*

Roy. Statist. Soc. **B 57**, 99-138.

[22] Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Institute Actuaries* **73**, 285-334.

[23] Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc.*, **B 49**, 223-239.

[24] Schervish, M. J. (1995). Theory of statistics. New York: Springer-Verlag.

[25] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461-464.

[26] Yuan, L. and Kesavan, H. (1997). Bayesian estimation of Shannon entropy. *Commun. Statist.-Theory Meth.*, **26**, 139-148.