BAYESIAN THEORY
Edited by José M. Bernardo and Adrian F. M. Smith

Chapter 4

Modelling

Summary

The relationship between beliefs about observable random quantities and their representation using conventional forms of statistical models is investigated. It is shown that judgements of exchangeability lead to representations that justify and clarify the use and interpretation of such familiar concepts as parameters, random samples, likelihoods and prior distributions. Beliefs which have certain additional invariance properties are shown to lead to representations involving familiar specific forms of parametric distributions, such as normals and exponentials. The concept of a sufficient statistic is introduced and related to representations involving the exponential family of distributions. Various forms of partial exchangeability judgements about data structures involving several samples, structured layouts, covariates and designed experiments are investigated, and links established with a number of other commonly used statistical models.

4.1 STATISTICAL MODELS

4.1.1 Beliefs and Models

The subjectivist, operationalist viewpoint has led us to the conclusion that, if we aspire to quantitative coherence, individual degrees of belief, expressed as probabilities, are inescapably the starting point for descriptions of uncertainty. There can

be no theories without theoreticians; no learning without learners; in general, no science without scientists. It follows that learning processes, whatever their particular concerns and fashions at any given point in time, are necessarily reasoning processes which take place in the minds of individuals. To be sure, the object of attention and interest may well be an assumed external, objective reality: but the actuality of the learning process consists in the evolution of individual, subjective beliefs about that reality. However, it is important to emphasise, as in our earlier discussion in Section 2.8, that the primitive and fundamental notions of *individual* preference and belief will typically provide the starting point for *interpersonal* communication and reporting processes. In what follows, both here, and more particularly in Chapter 5, we shall therefore often be concerned to identify and examine features of the individual learning process which relate to interpersonal issues, such as the conditions under which an approximate consensus of beliefs might occur in a population of individuals.

In Chapters 2 and 3, we established a very general foundational framework for the study of degrees of belief and their evolution in the light of new information. We now turn to the detailed development of these ideas for the broad class of problems of primary interest to statisticians; namely, those where the events of interest are defined explicitly in terms of *random quantities*, $x_1, \ldots, x_n$ (discrete or continuous, and possibly vector-valued) representing observed or experimental data.

In such cases, we shall assume that an individual's degrees of belief for events of interest are derived from the specification of a joint distribution function $P(x_1, \ldots, x_n)$, which we shall typically assume, without systematic reference to measure-theoretic niceties, to be representable in terms of a joint density function $p(x_1, \ldots, x_n)$ (to be understood as a mass function in the discrete case).

Of course, any such specification implicitly defines a number of other degrees of belief specifications of possible interest: for example, for $1 \leq m < n$,

$$p(x_1, \ldots, x_m) = \int p(x_1, \ldots, x_n) dx_{m+1} \ldots dx_n$$

provides the *marginal* joint density for $x_1, \ldots, x_m$, and

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = p(x_1, \ldots, x_n)/p(x_1, \ldots, x_m)$$

gives the joint density for the as yet unobserved $x_{m+1}, \ldots, x_n$, conditional on having observed $x_1 = x_1, \ldots, x_m = x_m$. Within the Bayesian framework, this latter conditional form is the key to "learning from experience".

We recall that, throughout, we shall use notation such as P and p in a *generic* sense, rather than as specifying particular functions. In particular, P may sometimes refer to an underlying probability measure, and sometimes refer to implied distribution functions, such as $P(x_1)$, $P(x_1, \ldots, x_n)$ or $P(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m)$. Similarly, we may write $p(x_1)$, $p(x_1, \ldots, x_n)$, etc. so that, for example,

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = p(x_1, \ldots, x_n)/p(x_1, \ldots, x_m)$$

simply indicates that the conditional density for $x_{m+1}, \ldots, x_n$ given $x_1, \ldots, x_m$ is given by the ratio of the specified joint densities. Such usage avoids notational proliferation, and the context will always ensure that there is no confusion of meaning.

Thus far, however, our discussion is rather "abstract". In actual applications we shall need to choose specific, concrete forms for joint distributions. This is clearly a somewhat daunting task, since direct contemplation and synthesis of the many complex marginal and conditional judgements implicit in such a specification are almost certainly beyond our capacity in all but very simple situations. We shall therefore need to examine rather closely this process of choosing a specific form of probability measure to represent degrees of belief.

Definition 4.1. (***Predictive probability model***). *A predictive model for a sequence of random quantities $x_1, x_2, \ldots$ is a probability measure P, which mathematically specifies the form of the joint belief distribution for any subset of $x_1, x_2, \ldots$.*

In some cases, we shall find that we are able to identify general types of belief structure which "pin down", in some sense, the mathematical representation strategy to be adopted. In other cases, this "formal" approach will not take us very far towards solving the representation problem and we shall have to fall back on rather more pragmatic modelling strategies.

At this stage, a word of warning is required. In much statistical writing, the starting point for formal analysis is the *assumption* of a mathematical model form, typically involving "unknown parameters", the main object of the study being to infer something about the values of these parameters. From our perspective, this is all somewhat premature and mysterious! We are seeking to represent degrees of belief about observables: nothing in our previous development justifies or gives any insight into the choice of particular "models", and thus far we have no way of attaching any operational meaning to the "parameters" which appear in conventional models. However, as we shall soon see, the subjectivist, operationalist approach *will* provide considerable insight into the nature and status of these conventional assumptions.

4.2 EXCHANGEABILITY AND RELATED CONCEPTS

4.2.1 Dependence and Independence

Consider a sequence of random quantities $x_1, x_2, \ldots$, and suppose that a predictive model is assumed which specifies that, for all n, the joint density can be written in

the form

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

so that the x_i are *independent* random quantities. It then follows straightforwardly that, for any $1 \leq m < n$,

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = p(x_{m+1}, \ldots, x_n).$$

so that *no learning from experience* can take place within this sequence of observations. In other works, past data provide us with no additional information about the possible outcomes of future observations in the sequence.

A predictive model specifying such an independence structure is clearly inappropriate in contexts where we believe that the successive accumulation of data will provide increasing information about future events. In such cases, the structure of the joint density $p(x_1, \ldots, x_n)$ must encapsulate some form of *dependence* among the individual random quantities. In general, however, there are a vast number of possible subjective assumptions about the form such dependencies might take and there can be no all-embracing theoretical discussion. Instead, what we can do is to concentrate on some particular simple forms of judgement about dependence structures which might correspond to actual judgements of individuals in certain situations.

There is no suggestion that the structures we are going to discuss in subsequent subsections have any special status, or *ought* to be adopted in most cases, or whatever. They simply represent forms of judgement which may often be felt to be appropriate and whose detailed analysis provides illuminating insight into the specification and interpretation of certain classes of predictive models.

4.2.2 Exchangeability and Partial Exchangeability

Suppose that, in thinking about $P(x_1, \ldots, x_n)$, his or her joint degree of belief distribution for a sequence of random quantities $x_1, \ldots, x_n$, an individual makes the judgement that the subscripts, the "labels" identifying the individual random quantities, are "uninformative", in the sense that he or she would specify all the marginal distributions for the individual random quantities identically, and similarly for all the marginal joint distributions for all possible pairs, triples, etc., of the random quantities. It is easy to see that this implies that the form of the joint distribution must be such that

$$P(x_1, \ldots, x_n) = P(x_{\pi(1)}, \ldots, x_{\pi(n)}),$$

for any possible permutation π of the subscripts $\{1, \ldots, n\}$. We formalise this notion of "symmetry" of beliefs for the individual random quantities as follows.

Definition 4.2. (***Finite exchangeability***). *The random quantities $x_1, \ldots, x_n$ are said to be judged (finitely) exchangeable under a probability measure P if the implied joint degree of belief distribution satisfies*

$$P(x_1, \ldots, x_n) = P(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

for all permutations π defined on the set $\{1, \ldots, n\}$. In terms of the corresponding density or mass function, the condition reduces to

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

Example 4.1. (***Tossing a thumb tack***). Consider a sequence of tosses of a standard metal drawing pin (or thumb tack), and let $x_i = 1$ if the pin lands point uppermost on the ith toss, $x_i = 0$ otherwise, $i = 1, \ldots, n$. If the tosses are performed in such a way that time order appears to be irrelevant and the conditions of the toss appear to be essentially held constant throughout, it would seem to be the case that, whatever precise *quantitative* form their beliefs take, most observers would judge the outcomes of the sequence of tosses $x_1, x_2, \ldots$ to be exchangeable in the above sense.

In general, the exchangeability assumption captures, for a subjectivist interested in belief distributions for observables, the essence of the idea of a so-called "random sample". This latter notion is, of course, of no direct use to us at this stage, since it (implicitly) involves the idea of "conditional independence, given the value of the underlying parameter", a meaningless phrase thus far within our framework.

The notion of exchangeability involves a judgement of *complete symmetry* among all the observables $x_1, \ldots, x_n$ under consideration. Clearly, in many situations this might be too restrictive an assumption, even though a partial judgement of symmetry is present.

Example 4.1. (***cont.***). Suppose that the sequence of tosses of a drawing pin are not all made with the same pin, but that the even and odd numbered tosses are made with different pins: an all metal one for the odd tosses; a plastic-coated one for the even tosses. Alternatively, suppose that the same pin were used throughout, but that the odd tosses are made by a different person, using a completely different tossing mechanism from that used for the even tosses. In such cases, many individuals would retain an exchangeable form of belief distribution within the sequences of odd and even tosses separately, but might be reluctant to make a judgement of symmetry for the combined sequence of tosses.

Example 4.2. ***(Laboratory measurements).*** Suppose that $x_1, x_2, \ldots$ are real-valued measurements of a physical or chemical property of a given substance, all made on the same sample with the same measurement procedure. Under such conditions, many individuals might judge the complete sequence of measurements to be exchangeable.

Suppose, however, that sequences of such measurements are combined from k different laboratories, the substance being identical but the measurement procedures varying from laboratory to laboratory. In this case, judgements of exchangeability for each laboratory sequence separately might be appropriate, whereas such a judgement for the combined sequence might not be.

Example 4.3. ***(Physiological responses).*** Suppose that $\{x_1, x_2, \ldots\}$ are real-valued measurements of a specific physiological response in human subjects when a particular drug is administered. If the drug is administered at more than one dose level and if there are both male and female subjects, spanning a wide age range, most individuals would be very reluctant to make a judgement of exchangeability for the entire sequence of results. However, within each combination of dose-level, sex and appropriately defined age-group, a judgement of exchangeability might be regarded as reasonable.

Judgements of the kind suggested in the above examples correspond to forms of *partial exchangeability*. Clearly, there are many possible forms of departure from overall judgements of exchangeability to those of partial exchangeability and so a formal definition of the term does not seem appropriate. In general, it simply signifies that there may be additional "labels" on the random quantities (for example, odd and even, or the identification of the tossing mechanism in Example 4.1) with exchangeable judgements made separately for each group of random quantities having the same additional labels. A detailed discussion of various possible forms of partial exchangeability will be given in Section 4.6.

We shall now return to the simple case of exchangeability and examine in detail the form of representation of $p(x_1, \ldots, x_n)$ which emerges in various special cases. As a preliminary, we shall generalise our previous definition of exchangeability to allow for "potentially infinite" sequences of random quantities. In practice, it should, at least in principle, always be possible to give an upper bound to the number of observables to be considered. However, specifying an actual upper bound may be somewhat difficult or arbitrary and so, for mathematical and descriptive purposes, it is convenient to be able to proceed as if we were contemplating an infinite sequence of potential observables. Of course, it will be important to establish that working within the infinite framework does not cause any fundamental conceptual distortion. These and related issues of finite versus infinite exchangeability will be considered in more detail in Section 4.7.1. For the time being, we shall concentrate on the "potentially infinite" case.

Definition 4.3. (***Infinite exchangeability***). *The infinite sequence of random quantities $x_1, x_2, \ldots$ is said to be judged (infinitely) exchangeable if every finite subsequence is judged exchangeable in the sense of Definition 4.2.*

One might be tempted to wonder whether *every* finite sequence of exchangeable random quantities could be embedded in or extended to an infinitely exchangeable sequence of similarly defined random quantities. However, this is certainly not the case as the following example shows.

Example 4.4. (***Non-extendible exchangeability***). Suppose that we define the three random quantities x_1, x_2, x_3 such that either $x_i = 1$ or $x_i = 0$, $i = 1, 2, 3$, with joint probability function given by

$$\begin{aligned} p(x_1 = 0, x_2 = 1, x_3 = 1) &= p(x_1 = 1, x_2 = 0, x_3 = 1) \\ &= p(x_1 = 1, x_2 = 1, x_3 = 0) \\ &= 1/3, \end{aligned}$$

with all other combinations of x_1, x_2, x_3 having probability zero, so that x_1, x_2, x_3 are clearly exchangeable. We shall now try to identify an x_4, taking only values 0 and 1, such that $x_1, \ldots, x_4$ are exchangeable. For this to be possible, we require, for example,

$$p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0) = p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1).$$

But

$$\begin{aligned} &p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0) \\ &\quad = p(x_1 = 0, x_2 = 1, x_3 = 1) - p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1) \\ &\quad = 1/3 - p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1) \\ &\quad = 1/3 - p(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 0), \end{aligned}$$

where

$$p(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 0) \leq p(x_1 = 1, x_2 = 1, x_3 = 1) = 0,$$

so that

$$p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0) = 1/3.$$

However, we also have

$$p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1) \leq p(x_1 = 0, x_2 = 0, x_3 = 1) = 0$$

and so

$$p(x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0) \neq p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1).$$

It follows that a finitely exchangeable sequence cannot even necessarily be embedded in a larger finitely exchangeable sequence, let alone an infinitely exchangeable sequence.

4.3 MODELS VIA EXCHANGEABILITY

4.3.1 The Bernoulli and Binomial Models

We consider first the case of an infinitely exchangeable sequence of 0–1 random quantities, $x_1, x_2, \ldots$ with $x_i = 0$ or $x_i = 1$, for all $i = 1, 2, \ldots$. Without loss of generality, we shall derive a representation result for the joint mass function, $p(x_1, \ldots, x_n)$, of the first n random quantities $x_1, \ldots, x_n$.

Proposition 4.1. (***Representation theorem for 0–1 random quantities***). *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0–1 random quantities with probability measure P, there exists a distribution function Q such that the joint mass function $p(x_1, \ldots, x_n)$ for $x_1, \ldots, x_n$ has the form*

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

where,

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

with $y_n = x_1 + \cdots + x_n$, and $\theta = \lim_{n \to \infty} y_n/n$.

Proof. (De Finetti, 1930, 1937/1964; here we follow closely the proof given by Heath and Sudderth, 1976; see also Barlow, 1991). Suppose $x_1 + \cdots + x_n = y_n$, then, by exchangeability, for any $0 \leq y_n \leq n$,

$$p(x_1 + \cdots + x_n = y_n) = \binom{n}{y_n} p(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

for any permutation π of $\{1, \ldots, n\}$ such that $x_{\pi(1)} + \cdots + x_{\pi(n)} = y_n$. Moreover, for arbitrary $N \geq n \geq y_n \geq 0$, and with the summations below taken over the range $y_N = y_n$ to $y_N = N - (n - y_n)$, we see that

$$\begin{aligned}
&p(x_1 + \cdots + x_n = y_n) \\
&= \sum p(x_1 + \cdots + x_n = y_n \mid x_1 + \cdots + x_N = y_N)\, p(x_1 + \cdots + x_N = y_N). \\
&= \sum \binom{N}{n}^{-1} \binom{y_N}{y_n} \binom{N - y_N}{n - y_n} p(x_1 + \cdots + x_N = y_N), \quad 0 \leq y_n \leq n \leq N. \\
&= \binom{n}{y_n} \sum \frac{(y_N)_{y_n} (N - y_N)_{n - y_n}}{(N)_n} p(x_1 + \cdots + x_N = y_N).
\end{aligned}$$

where $(y_N)_{y_n} = y_N(y_N - 1) \cdots [y_N - (y_n - 1)]$, etc. (Intuitively, we can imagine sampling n items without replacement from an urn of N items containing y_N 1's and $N - y_N$ 0's, corresponding to the hypergeometric distribution of Section 3.2.2.)

If we now define $Q_N(\theta)$ on $\Re$ to be the step function which is 0 for $\theta < 0$ and has jumps of $p(x_1 + \cdots + x_N = y_N)$ at $\theta = y_N/N, y_N = 0, \ldots, N$, we see that

$$p(x_1 + \cdots + x_n = y_n) = \binom{n}{y_n} \int_0^1 \frac{(\theta N)_{y_n}[(1-\theta)N]_{n-y_n}}{(N)_n} \, dQ_N(\theta).$$

As $N \to \infty$,

$$\frac{(\theta N)_{y_n}[(1-\theta)N]_{n-y_n}}{(N)_n} \to \theta^{y_n}(1-\theta)^{n-y_n}$$

uniformly in θ. Moreover, by Helly's theorem (see, for example, Section 3.2.3 and Ash, 1972, Section 8.2), there exists a subsequence $Q_{N_1}, Q_{N_2}, \ldots$ such that

$$\lim_{N_j \to \infty} Q_{N_j} = Q,$$

where Q is a distribution function. The result follows. ◁

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

(i) the x_i are judged to be independent, Bernoulli random quantities (see Section 3.2.2) conditional on a random quantity θ;
(ii) θ is itself assigned a probability distribution Q;
(iii) 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 relative frequency of 1's".

In more conventional notation and language, it is as if, conditional on θ, $x_1, \ldots, x_n$ are a *random sample* from a Bernoulli distribution with *parameter* θ, generating a parametrised *joint sampling distribution*

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

where the parameter is assigned a *prior distribution* $Q(\theta)$. The operational content of this prior distribution derives from the fact that it is *as if* we are assessing beliefs about what we would anticipate observing as the *limiting relative frequency* from a "very large number" of observations. Thought of as a function of θ, we shall refer to the joint sampling distribution as the *likelihood* function.

In terms of Definition 4.1, the assumption of exchangeability for the infinite sequence of 0–1 random quantities $x_1, x_2, \ldots$ places a strict limitation on the family of probability measures P which can serve as predictive probability models for the sequence. Any such P must correspond to the mixture form given in Proposition 4.1, for some choice of prior distribution $Q(\theta)$. As we range over all possible choices of this latter distribution, we generate all possible predictive

probability models compatible with the assumption of infinite exchangeability for the 0–1 random quantities.

Thus, "at a stroke", we establish 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; the latter, and its associated prior distribution, acquire an operational interpretation in terms of a limiting average of observables (in this case a limiting frequency).

In many applications involving 0–1 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 of $p(x_1 + \cdots + x_n = y_n)$ is straightforwardly obtained from Proposition 4.1.

Corollary 1. *Given the conditions of Proposition 4.1,*

$$p(x_1 + \cdots + x_n = y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1-\theta)^{n-y_n} dQ(\theta).$$

Proof. This follows immediately from Proposition 4.1 and the fact that

$$p(x_1 + \cdots + x_n = y_n) = \binom{n}{y_n} p(x_1, \ldots, x_n)$$

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

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

The formal learning process for models such as this will be developed systematically and generally in Chapter 5. However, this simple example provides considerable insight into the learning process, showing how, in a sense, the key step is a straightforward consequence of the representation theorem.

Corollary 2. *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0–1 random quantities with probability measure P, the conditional probability function $p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m)$, for $x_{m+1}, \ldots, x_n$ given $x_1, \ldots, x_m$, has the form*

$$\int_0^1 \prod_{i=m+1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} dQ(\theta \mid x_1, \ldots, x_m), \quad 1 \leq m < n.$$

where

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

and

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

Proof. Clearly,

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_m)},$$

and the result follows by applying Proposition 4.1 to both $p(x_1, \ldots, x_n)$ and $p(x_1, \ldots, x_m)$ and rearranging the resulting expression. ◁

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, via *Bayes' theorem*, into the *posterior* distribution $Q(\theta \mid x_1, \ldots, x_m)$.

The conditional probability function $p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m)$ is called the (*conditional, or posterior*) *predictive* probability function for $x_{m+1}, \ldots, x_n$ given $x_1, \ldots, x_m$, and this, of course, also provides the basis for deriving the conditional predictive distribution of any other random quantity defined in terms of the future observations. For example, given $x_1, \ldots, x_m$, the predictive probability function $p(y_{n-m} \mid x_1, \ldots, x_m)$ for y_{n-m}, i.e., the total number of 1's in $x_{m+1}, \ldots, x_n$, has the form

$$\int_0^1 \binom{n-m}{y_{n-m}} \theta^{y_{n-m}}(1-\theta)^{(n-m)-y_{n-m}} dQ(\theta \mid x_1, \ldots, x_m).$$

A particularly important random quantity defined in terms of future observations is the frequency of 1's in a large sample. But, by Proposition 4.1 and its Corollary 2,

$$\lim_{(n-m)\to\infty} P\left(\frac{y_{n-m}}{(n-m)} \leq \theta \middle| x_1, \ldots, x_m\right) = Q(\theta \mid x_1, \ldots, x_m).$$

Thus, *a posterior distribution for a parameter is seen to be a limiting case of a posterior (conditional) predictive distribution for an observable.*

4.3.2 The Multinomial Model

An alternative way of viewing the 0–1 random quantities discussed in Section 3.1 is as defining category membership (given two exclusive and exhaustive categories), in the sense that $x_i = 1$ signifies that the ith observation belongs to category 1 and $x_i = 0$ signifies membership of category 2. We can extend this idea in an obvious way by considering k-dimensional random vectors $\boldsymbol{x}_i$ whose jth component, x_{ij}, takes the value 1 to indicate membership of the jth of $k+1$ categories. At most one of the k components can take the value 1; if they all take the value 0 this signifies membership of the $(k+1)$th category. In what follows, we shall refer to such $\boldsymbol{x}_i$ as "0–1 random vectors". If $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is an infinitely exchangeable sequence of 0–1 random vectors, we can extend Proposition 4.1 in an obvious way.

Proposition 4.2. (***Representation theorem for 0–1 random vectors***). *If $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is an infinitely exchangeable sequence of 0–1 random vectors with probability measure P, there exists a distribution function Q such that the joint mass function $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ has the form*

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \int_{\Theta^*} \prod_{i=1}^{n} \theta_1^{x_{i1}} \theta_2^{x_{i2}} \cdots \theta_k^{x_{ik}} \left(1 - \sum_{j=1}^{k} \theta_j\right)^{1 - \sum x_{ij}} dQ(\boldsymbol{\theta}),$$

where

$$\Theta^* = \left\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k): \quad 0 \leq \theta_i \leq 1, \ \sum_{i=1}^{k} \theta_i \leq 1 \right\}$$

and

$$Q(\boldsymbol{\theta}) = \lim_{n \to \infty} P\left[\left(\overline{x}_{1n} \leq \theta_1, \ldots\right) \cup \ldots \cup \left(\overline{x}_{kn} \leq \theta_k\right) \right],$$

with $\overline{x}_{in} = n^{-1}(x_{1i} + \cdots + x_{ni})$, and $\theta_i = \lim_{n \to \infty} \overline{x}_{in}$.

Proof. This is a straightforward, albeit algebraically cumbersome, generalisation of the proof of Proposition 4.1. ◁

As in the previous case, we are often most interested in the summary random vector $\boldsymbol{y}_n = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$ whose jth component y_{nj} is the random quantity corresponding to the total number of occurrences of category j in the n observations. We shall give the representation of $p(\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n = \boldsymbol{y}_n) = p(y_{n1}, \ldots, y_{nk})$, generalising Corollary 1 to Proposition 4.1, and then comment on the interpretation of these results.

Corollary. *Given the conditions of Proposition 4.2, the joint mass function* $p(y_{n1}, \ldots, y_{nk})$ *may be represented as*

$$\int_{\Theta^*} \binom{n}{y_{n1} \cdots y_{nk}} [\theta_1^{y_{n1}} \theta_2^{y_{n2}} \cdots \theta_k^{y_{nk}}](1 - \Sigma\theta_i)^{n-\Sigma y_{ni}} dQ(\boldsymbol{\theta})$$

where

$$\binom{n}{y_{n1} \cdots y_{nk}} = \frac{n!}{y_{n1}! y_{n2}! \cdots y_{nk}!(n - \Sigma y_{ni})!},$$

Proof. This follows immediately from the generalisation of the argument used in proving Corollary 1 to Proposition 4.1. ◁

Thus, we see in Proposition 4.2 that it is *as if* we have a likelihood corresponding to the joint sampling distribution of a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where each $\boldsymbol{x}_i$ has a multinomial distribution with probability function $\text{Mu}_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}, 1)$, together with a prior distribution Q over the multinomial parameter $\boldsymbol{\theta}$, where the components θ_j of the latter can be thought of as the limiting relative frequency of membership of the jth category. In the corollary, it is *as if* we assume a multinomial likelihood, $\text{Mu}_k(\boldsymbol{y}_n \mid \boldsymbol{\theta}, n)$, with a prior $Q(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$.

4.3.3 The General Model

We now consider the case of an infinitely exchangeable sequence of real-valued random quantities $x_1, x_2, \ldots$. As one might expect, the mathematical technicalities of establishing a representation theorem in the real-valued case are somewhat more complicated than in the 0–1 cases, and a rigorous treatment involves the use of measure-theoretic tools beyond the general mathematical level at which this volume is aimed. For this reason, we shall content ourselves with providing an *outline proof* of a form of the representation theorem, having no pretence at mathematical rigour but, hopefully, providing some intuitive insight into the result, as well as the key ideas underlying a form of proper proof.

Proposition 4.3. (***General representation theorem***).
If $x_1, x_2, \ldots$, *is an infinitely exchangeable sequence of real-valued random quantities with probability measure* P, *there exists a probability measure* Q *over* $\Im$, *the space of all distribution functions on* $\Re$, *such that the joint distribution function of* $x_1, \ldots, x_n$ *has the form*

$$P(x_1, \ldots, x_n) = \int_{\Im} \prod_{i=1}^{n} F(x_i) dQ(F),$$

where

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

and F_n *is the empirical distribution function defined by* $x_1, \ldots, x_n$.

Outline proof. (See Chow and Teicher, 1978/1988). Since

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(x_i \le x)}$$

we have, by exchangeability,

$$\mathrm{E}(F_n(x) - F_N(x))^2 = \frac{|N-n|}{Nn}\Big\{P(x_1 < x) - P[(x_1 < x) \cap (x_2 < x)]\Big\}.$$

To see this, writing I_i in place of $I_{(x_i \le x)}$ and noting that $I_i^2 = I_i$, we have

$$\begin{aligned}\left[F_n(x) - F_N(x)\right]^2 = &\left(\frac{1}{n^2}\sum_{i=1}^{n} - \frac{1}{N^2}\sum_{i=1}^{N}\frac{2}{nN}\sum_{i=1}^{n}\right)(I_i) \\ &+ 2\left(\frac{1}{n^2}\sum_{j}\sum_{i<j}^{n} - N^2\sum_{j}\sum_{i<j}^{N}\frac{1}{nN}\sum_{j}\sum_{i<j}^{n}\right)(I_iI_j).\end{aligned}$$

Note also that $\mathrm{E}(I_i) = P(x_1 < x)$ and $\mathrm{E}(I_iI_j) = P[(x_1 < x) \cap (x_2 < x)]$, for all i, j, by exchangeability. A straightforward count of the numbers of terms involved in the summations then gives the required result.

The right-hand side tends to zero as $N, n \to \infty$, and hence the random quantity $F_n(x)$ tends in probability to some random quantity, $F(x)$, say, which implies that

$$\prod_{j=1}^{n} F_N(x_j) \to \prod_{j=1}^{n} F(x_j) \qquad (*)$$

in probability as $N \to \infty$, for fixed n.

Suppose we now let $\alpha_1, \ldots, \alpha_n$ denote positive integers and set

$$A = \{\alpha = (\alpha_1, \ldots, \alpha_n): \quad 1 \le \alpha_i \le N \text{ for } 1 \le i \le n\}$$

and

$$A^* = I(\alpha) = I[(x_{\alpha_1} \le x_1) \cap \cdots \cap (x_{\alpha_n} \le x_n)].$$

For $N > n$, it then follows that

$$\begin{aligned}\prod_{j=1}^{n} F_N(x_j) &= N^{-n}\prod_{j=1}^{n}\sum_{i=1}^{N} I_{(x_i \le x_j)} = N^{-n}\sum_{\alpha \in A} I(\alpha) \\ &= N^{-n}\left(\sum_{\alpha \in A - A^*} + \sum_{\alpha \in A^*}\right) I(\alpha).\end{aligned}$$

However, as $N \to \infty$,

$$N^{-n} \sum_{\alpha \in A-A^*} I(\alpha) \leq N^{-n} \sum_{\alpha \in A-A^*} 1 = [N^n - N(N-1)\cdots(N-n+1)]/N^n \to 0$$

so that,

$$\prod_{j=1}^{n} F_N(x_j) \approx N^{-n} \sum_{\alpha \in A^*} I(\alpha)$$

But, by exchangeability,

$$\begin{aligned} \int I(\alpha)\, dP &= \int I_{[(x \leq x_1) \cap \cdots \cap (x \leq x_n)]}\, dP \\ &= P[(x \leq x_1) \cap \cdots \cap (x \leq x_n)] = P(x_1, \ldots, x_n) \end{aligned}$$

and so

$$\int \prod_{i=1}^{n} F_N(x_j) dP \approx [N \cdots (N-n+1)/N^n] P[(x \leq x_1) \cap \cdots \cap (x \leq x_n)].$$

Recalling (*), we see that, as $N \to \infty$,

$$\int \prod_{j=1}^{n} F(x_j)\, dQ(F) \approx P(x_1, \ldots, x_n)$$

where $Q(F) = \lim_{N \to \infty} P(F_N)$. ◁

The general form of representation for real-valued exchangeable random quantities is therefore *as if* we have independent observations $x_1, \ldots, x_n$ conditional on F, an unknown (i.e., random) distribution function (which plays the role of an infinite-dimensional "parameter" in this case), with a belief distribution Q for F, having the operational interpretation of "what we believe the empirical distribution function would look like for a large sample".

The structure of the learning process for a general exchangeable sequence of real-valued random quantities, with the distribution function representation given in Proposition 4.3, cannot easily be described explicitly. In what follows, we shall therefore find it convenient to restrict attention to those cases where a corresponding representation holds in terms of density functions, labelled by a finite-dimensional parameter, θ, say, rather than the infinite-dimensional label, F. For ease of reference, we present this representation as a corollary to Proposition 4.3.

Corollary 1. *Assuming the required densities to exist, under the conditions of Proposition 4.3 the joint density of $x_1, \ldots, x_n$ has the form*

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})\, dQ(\boldsymbol{\theta}),$$

with $p(\cdot \mid \boldsymbol{\theta})$ denoting the density function corresponding to the "unknown parameter" $\boldsymbol{\theta} \in \Theta$.

The role of Bayes' theorem in the learning process is now easily identified.

Corollary 2. *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of real-valued random quantities admitting a density representation as in Corollary 1, then*

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = \int_\Theta \prod_{i=m+1}^{n} p(x_i \mid \boldsymbol{\theta})\, dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m)$$

where

$$dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m) = \frac{\prod_{i=1}^{m} p(x_i \mid \boldsymbol{\theta})\, dQ(\boldsymbol{\theta})}{\int_\Theta \prod_{i=1}^{m} p(x_i \mid \boldsymbol{\theta})\, dQ(\boldsymbol{\theta})}.$$

Proof. This follows immediately on writing

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_m)},$$

applying the density representation form to both $p(x_1, \ldots, x_n)$ and $p(x_1, \ldots, x_m)$, and rearranging the resulting expression. ◁

The technical discussion in this section has centred on exchangeable sequences, $x_1, x_2, \ldots$, of real-valued random quantities. In fact, everything carries over in an obviously analogous manner to the case of exchangeable sequences $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, with $\boldsymbol{x}_i \in \Re^k$. All that happens, in effect, is that the distribution functions and densities referred to in Proposition 4.3 and its corollaries become the joint distribution functions and densities for the k components of the $\boldsymbol{x}_i$. To avoid tedious distinctions between $x \in \Re$ and $\boldsymbol{x} \in \Re^k$, in subsequent developments we shall often just write $x \in X$. In cases where the distinction between $k = 1$ and $k > 1$ matters, it will be clear from the context what is intended.

In Section 4.8.1, we shall give detailed references to the literature on representation theorems for exchangeable sequences, including far-reaching generalisations of the 0–1 and real-valued cases. However, even the simple cases we have presented already provide, from the subjectivist perspective, a deeply satisfying clarification of such fundamental notions as *models, parameters, conditional independence* and the relationship between *beliefs* and *limiting frequencies*.

In terms of Definition 4.1, the assumption of exchangeability for the real-valued random quantities $x_1, x_2, \ldots$ again places (as in the 0–1 case) a limitation on the family of probability measures P which can serve as predictive probability models. In this case, however, in the context of the general form of representation given in Proposition 4.3, the "parameter", F, underlying the conditional independence structure within the mixture is a random distribution function, so that the "parameter" is, in effect, *infinite dimensional*, and the family of coherent predictive probability models is generated by ranging through all possible prior distributions $Q(F)$. The mathematical form of the required representation is well-defined, but the practical task of translating actual beliefs about real-valued random quantities into the required mathematical form of a measure over a function space seems, to say the least, a somewhat daunting prospect. It is interesting therefore to see whether there exist more complex formal structures of belief, imposing further symmetries or structure beyond simple exchangeability, which lead to more specific and "familiar" model representations. In particular, it is of interest to identify situations in which exchangeability leads to a mixture of conditional independence structures which are defined in terms of a *finite dimensional* parameter so that the more explicit forms given in the corollaries to Proposition 4.3 can be invoked. Given the interpretation of the components of such a parameter as strong law limits of simple sequences of functions of the observations, the specification of Q, and hence of the complete predictive probability model P, then becomes a much less daunting task.

4.4 MODELS VIA INVARIANCE

4.4.1 The Normal Model

Suppose that in addition to judging an infinite sequence of real-valued random quantities $x_1, x_2, \ldots$ to be exchangeable, we consider the possibility of further judgements of invariance, perhaps relating to the "geometry" of the space in which a finite subset of observations, $x_1, \ldots, x_n$, say, lie. The following definitions describe two such possible judgements of invariance. As with exchangeability, there is no claim that such judgements have any a priori special status. They are intended, simply, as possible forms of judgement that *might* be made, and whose consequences might be interesting to explore.

Definition 4.4. (***Spherical symmetry***). *A sequence of random quantities $x_1, \ldots, x_n$ is said to have spherical symmetry under a predictive probability model P if the latter defines the distributions of $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{A}\boldsymbol{x}$ to be identical, for any (orthogonal) $n \times n$ matrix $\boldsymbol{A}$ such that $\boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{I}$.*

This definition encapsulates a judgement of rotational symmetry, in the sense that, although measurements happened to have been expressed in terms of a particular coordinate system (yielding $x_1, \ldots, x_n$), our quantitative beliefs would not change if they had been expressed in a rotated coordinate system. Since rotational invariance fixes "distances" from the origin, this is equivalent to a judgement of identical beliefs for all outcomes of $x_1, \ldots, x_n$ leading to the same value of $x_1^2 + \cdots + x_n^2$.

The next result states that if we make the judgement of spherical symmetry (which in turn implies a judgement of exchangeability, since permutation is a special case of orthogonal transformation), the general mixture representation given in Proposition 4.3 assumes a much more concrete and familiar form.

Proposition 4.4. (***Representation theorem under spherical symmetry***). *If $x_1, x_2, \ldots$ is an infinite sequence of real-valued random quantities with probability measure P, and if, for any n, $\{x_1, \ldots, x_n\}$ have spherical symmetry, there exists a distribution function Q on $\Re^+$ such that the joint distribution function of $x_1, \ldots, x_n$ has the form*

$$P(x_1, \ldots, x_n) = \int_{\Re^+} \prod_{i=1}^{n} \Phi(\lambda^{1/2} x_i) dQ(\lambda),$$

where Φ is the standard normal distribution function and

$$Q(\lambda) = \lim_{n \to \infty} P(s_n^{-2} \leq \lambda),$$

with $s_n^2 = n^{-1}(x_1^2 + \cdots + x_n^2)$, and $\lambda^{-1} = \lim_{n \to \infty} s_n^2$.

Proof. See, for example, Freedman (1963a) and Kingman (1972); details are omitted here, since the proof of a generalisation of this result will be given in full in Proposition 4.5. ◁

The form of representation obtained in Proposition 4.4 tells us that the judgement of spherical symmetry restricts the set of coherent predictive probability models to those which are generated by acting *as if*:

(i) observations are conditionally independent *normal* random quantities, given the random quantity λ (which, as a "labelling parameter", corresponds to the precision; i.e., the reciprocal of the variance);
(ii) λ is itself assigned a distribution Q;
(iii) by the strong law of large numbers, $\lambda^{-1} = \lim_{n \to \infty} s_n^2$, so that Q may be interpreted as "beliefs about the reciprocal of the limiting mean sum of squares of the observations".

For related work see Dawid (1977, 1978). To obtain a justification for the usual normal specification, with "unknown mean and precision", we need to generalise the above discussion slightly.

We note first that the judgement of spherical symmetry implicitly attaches a special significance to the *origin* of the coordinate system, since it is equivalent to a judgement of invariance in terms of distance from the origin. In general, however, if we were to feel able to make a judgement of spherical symmetry, it would typically only be *relative* to an "origin" defined in terms of the "centre" of the random quantities under consideration. This motivates the following definition.

Definition 4.5. (***Centred spherical symmetry***). *A sequence of random quantities $x_1, \ldots, x_n$ is said to have centred spherical symmetry if the random quantities $x_1 - \bar{x}_n, \ldots, x_n - \bar{x}_n$ have spherical symmetry, where $\bar{x}_n = n^{-1} \sum x_i$. This is equivalent to a judgement of identical beliefs for all outcomes of $x_1, \ldots, x_n$ leading to the same value of $(x_1 - \bar{x}_n)^2 + \cdots + (x_n - \bar{x}_n)^2$.*

Proposition 4.5. (***Representation under centred spherical symmetry***). *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of real-valued random quantities with probability measure P, and if, for any n, $\{x_1, \ldots x_n\}$ have centred spherical symmetry, then there exists a distribution function Q on $\Re \times \Re^+$ such that the joint distribution of $x_1, \ldots, x_n$ has the form*

$$P(x_1, \ldots, x_n) = \int_{\Re \times \Re^+} \prod_{i=1}^{n} \Phi[\lambda^{1/2}(x_i - \mu)] dQ(\mu, \lambda),$$

where Φ is the standard normal distribution function and

$$Q(\mu, \lambda) = \lim_{n \to \infty} P\big[(\bar{x}_n \leq \mu) \cap (s_n^{-2} \leq \lambda)\big],$$

with $\bar{x}_n = n^{-1}(x_1 + \cdots + x_n)$, $s_n^2 = n^{-1}\big[(x_1 - \bar{x}_n)^2 + \cdots + (x_n - \bar{x}_n)^2\big]$, $\mu = \lim_{n \to \infty} \bar{x}_n$, and $\lambda^{-1} = \lim_{n \to \infty} s_n^2$.

Proof. (Smith, 1981). Since the sequence $x_1, x_2, \ldots$ is exchangeable, by Proposition 4.3 there exists a random distribution function F such that, conditional on F, the random quantities $x_1, \ldots, x_n$, for any n, are independent. There is therefore a random characteristic function, ϕ, corresponding to F, such that

$$E\left[\exp\left(i \sum_{j=1}^{n} t_j x_j\right) \middle| F\right] = \prod_{j=1}^{n} \phi(t_j)$$

and hence

$$E\left[\exp\left(i \sum_{j=1}^{n} t_j x_j\right)\right] = E\left[\prod_{j=1}^{n} \phi(t_j)\right].$$

If we now define $y_j = x_j - \bar{x}_n$, $j = 1, \ldots, n$, it follows that

$$E\left[\exp\left(i\sum_{j=1}^{n} s_j y_j\right)\right] = E\left[\prod_{j=1}^{n}\phi(s_j)\right] \qquad (*)$$

for all real $s_1, \ldots, s_n$ such that $s_1 + \cdots + s_n = 0$. Since $y_1, \ldots, y_n$ are spherically symmetric, both sides of this latter equality depend only on $s_1^2 + \cdots + s_n^2$.

Recalling that $\phi(-t) = \bar{\phi}(t)$, the complex conjugate, and that $\phi(0) = 1$, it follows that, for any real u and v,

$$\begin{aligned} E\{|\phi(u+v)\phi(u-v) - \phi^2(u)\phi(v)\phi(-v)|^2\} \\ = E\{\phi(u+v)\phi(u-v)\phi(-u-v)\phi(v-u)\} \\ - E\{\phi(u+v)\phi(u-v)\phi^2(-u)\phi(-v)\phi(v)\} \\ - E\{\phi(-u-v)\phi(v-u)\phi^2(u)\phi(v)\phi(-v)\} \\ + E\{\phi^2(u)\phi^2(v)\phi^2(-v)\phi^2(-u)\}. \end{aligned}$$

where all four terms in this expression are of the form of the right-hand side of (*) with $n = 8$, $s_1 + \cdots + s_8 = 0$ and $s_1^2 + \cdots + s_8^2 = 4(u^2 + v^2)$. All the four terms are therefore equal, so that the overall expression is zero. This implies that, almost surely with respect to the probability measure P, ϕ satisfies the functional equation

$$\phi(u+v)\phi(u-v) = \phi^2(u)\phi(v)\phi(-v)$$

for all real u and v. This can be rewritten in the form

$$\Psi_1(u+v) + \Psi_2(u-v) = A(u) + B(v),$$

where $\Psi_1(t) = \Psi_2(t) = \log\phi(t)$, and where $A(u) = 2\log\phi(u)$ and $B(v) = \log[\phi(v)\phi(-v)]$; it follows that $\log\phi(t)$ is a quadratic in t (see, for example, Kagan, Linnik and Rao, 1973, Lemma 1.5.1). Again using $\phi(-t) = \bar{\phi}(t)$, $\phi(0) = 1$, we see that, for this quadratic, the constant coefficient must be zero, the linear coefficient purely imaginary and the quadratic coefficient real and non-positive. This establishes that the random characteristic function ϕ takes the form

$$\phi(t) = \exp\left\{i\mu t - \frac{1}{2}\frac{t^2}{\lambda}\right\}$$

for some random quantities $\mu \in \Re$, $\lambda \in \Re^+$.

If we now define a random quantity z by

$$z = \exp\left(i\sum_{j=1}^{n} t_j x_j\right).$$

then, by iterated expectation, we have

$$E(z \mid \mu, \lambda) = E[E(z \mid F) \mid \mu, \lambda] = E\left[\prod_{j=1}^{n} \phi(t_j) \middle| \mu, \lambda\right]$$

so that

$$E\left[\exp\left(i \sum_{j=1}^{n} t_j x_j\right) \middle| \mu, \lambda\right] = \prod_{j=1}^{n} \exp\left(i\mu t_j - \frac{1}{2} \frac{t_j^2}{\lambda}\right).$$

This establishes that, conditional on μ and λ, $x_1, \ldots, x_n$ are independent normally distributed random quantities, each with mean μ and precision λ. The mixing distribution in the general representation theorem reduces therefore to a joint distribution over μ and λ. But, by the strong law of large numbers,

$$\lim_{n \to \infty} \frac{x_1 + \cdots + x_n}{n} = \mu,$$

$$\lim_{n \to \infty} \frac{(x_1 - \bar{x}_n)^2 + \cdots + (x_n - \bar{x}_n)^2}{n} = \frac{1}{\lambda},$$

and the result follows. ◁

We see, therefore, that the combined judgements of exchangeability and centred spherical symmetry restrict the set of coherent predictive probability models to those which, expressed in conventional terminology, correspond to acting *as if*:

(i) we have a *random sample* from a *normal distribution* with *unknown mean and precision* parameters, μ and λ, generating a *likelihood*

$$p(x_1, \ldots, x_n \mid \mu, \lambda) = \prod_{i=1}^{n} N(x_i \mid \mu, \lambda);$$

(ii) we have a joint *prior distribution* $Q(\mu, \lambda)$ for the unknown parameters, μ and λ, which can be given an operational interpretation as "beliefs about the sample mean and reciprocal sample variance which would result from a large number of observations".

4.4.2 The Multivariate Normal Model

Suppose now that we have an infinitely exchangeable sequence of random vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ taking values in $\Re^k, k \geq 2$, and that, in addition, we judge, for all n and for all $\boldsymbol{c} \in \Re^k$, that the random quantities $\boldsymbol{c}^t\boldsymbol{x}_1, \ldots, \boldsymbol{c}^t\boldsymbol{x}_n$ have centred spherical symmetry. The next result then provides a multivariate generalisation of Proposition 4.5.

Proposition 4.6. (***Multivariate representation theorem under centred spherical symmetry***). *If $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is an infinitely exchangeable sequence of random vectors taking values in $\Re^k$, with probability measure P, such that, for any n and $\boldsymbol{c} \in \Re^k$, the random quantities $\boldsymbol{c}'\boldsymbol{x}_1, \ldots, \boldsymbol{c}'\boldsymbol{x}_n$ have centred spherical symmetry, the structure of evaluations under P of probabilities of events defined by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is as if the latter were independent, multivariate normally distributed random vectors, conditional on a random mean vector $\boldsymbol{\mu}$ and a random precision matrix $\boldsymbol{\lambda}$, with a distribution over $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ induced by P, where*

$$\boldsymbol{\mu} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i, \quad \boldsymbol{\lambda}^{-1} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)'.$$

Proof. Defining $y_j = \boldsymbol{c}'\boldsymbol{x}_j$, $j = 1, \ldots, n$, we see that the random quantities $y_1, \ldots, y_n$ have centred spherical symmetry and so, by Proposition 4.5, there exist $\mu = \mu(\boldsymbol{c})$ and $\lambda = \lambda(\boldsymbol{c})$ such that, for all $t_j \in \Re$, $j = 1, \ldots, n$.

$$E\left[\exp\left(i \sum_{j=1}^{n} t_j y_j\right) \middle| \mu, \lambda\right] = \prod_{j=1}^{n} \exp\left(i\mu t_j - \frac{1}{2} \frac{t_j^2}{\lambda}\right).$$

where

$$\mu = \mu(\boldsymbol{c}) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} y_j, \quad \lambda^{-1} = \lambda^{-1}(\boldsymbol{c}) = \lim_{n\to\infty} \sum_{j=1}^{n} (y_j - y_n)^2.$$

But

$$\mu = \mu(\boldsymbol{c}) = \boldsymbol{c}' \lim_{n\to\infty} \left(\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j\right) = \boldsymbol{c}'\boldsymbol{\mu}$$

and

$$\lambda^{-1} = \lambda^{-1}(\boldsymbol{c}) = \boldsymbol{c}' \lim_{n\to\infty} \left[\frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_j - \bar{\boldsymbol{x}}_n)'\right] \boldsymbol{c} = \boldsymbol{c}'\boldsymbol{\lambda}^{-1}\boldsymbol{c}.$$

so that

$$E\left[\exp\left(i\boldsymbol{c}' \sum_{j=1}^{n} t_j \boldsymbol{x}_j\right) \middle| \boldsymbol{\mu}, \boldsymbol{\lambda}\right] = \prod_{j=1}^{n} \exp[i\boldsymbol{c}'\boldsymbol{\mu} t_j - \tfrac{1}{2}(\boldsymbol{c}'\boldsymbol{\lambda}^{-1}\boldsymbol{c} t_j^2)].$$

for all $\boldsymbol{c} \in \Re^k, t_j \in \Re, j = 1, \ldots, n$. It follows that, for all $\boldsymbol{t}_j \in \Re^k, j = 1, \ldots, n$.

$$E\left[\exp\left(i \sum_{j=1}^{n} \boldsymbol{t}_j' \boldsymbol{x}_j\right) \middle| \boldsymbol{\mu}, \boldsymbol{\lambda}\right] = \prod_{j=1}^{n} \exp[i\boldsymbol{\mu}'\boldsymbol{t}_j - \tfrac{1}{2}(\boldsymbol{t}_j'\boldsymbol{\lambda}^{-1}\boldsymbol{t}_j)]$$

so that, conditional on $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent multivariate normal random quantities each with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\lambda}$. ◁

4.4.3 The Exponential Model

Suppose $x_1, x_2, \ldots$ is judged to be an infinitely exchangeable sequence of positive real-valued random quantities. In particular, we note that this implies, for any pair x_i, x_j, an identity of beliefs for any events in the positive quadrant which are symmetrically placed with respect to the 45° line through the origin.

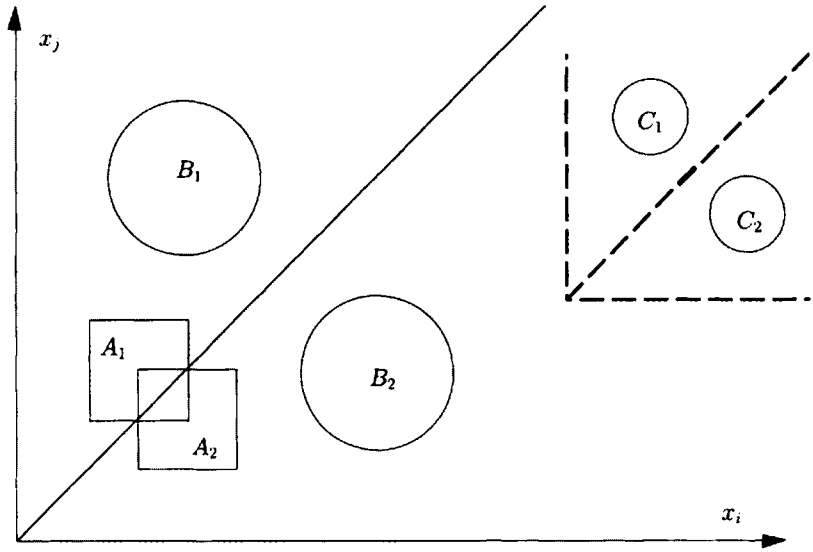

Figure 4.1 *A_1, A_2, B_1, B_2 reflections in 45° line. C_1, C_2 reflections in (dashed) 45° line*

Thus, for example, in Figure 4.1, the probabilities assigned to A_1 and A_2, B_1 and B_2, respectively, must be equal, for any $i \neq j$. In general, however, the assumption of exchangeability would not imply that events such as C_1 and C_2 have equal probabilities, even though they are symmetrically placed with respect to a 45° line (but not the one through the origin).

It is interesting to ask under what circumstances an individual *might* judge events such as C_1, C_2 to have equal probabilities. The answer is suggested by the additional (dashed) lines in the figure. *If* we added to the assumption of exchangeability the judgement that the "origins" of the x_i and x_j axes are "irrelevant", so far as probability judgements are concerned, then the probabilities of events such as C_1 and C_2 *would* be judged equal. In perhaps more familiar terms, this would be as though, when making judgements about events in the positive quadrant, an individual's judgement exhibited a form of "lack of memory" property with respect to the origin. If such a judgement is assumed to hold for all subsets of n (rather than just two) random quantities, the resulting representation is as follows.

Proposition 4.7. (***Continuous representation under origin invariance***). *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of positive real-valued random quantities with probability measure P, such that, for all n, and any event A in $\Re^+ \times \ldots \times \Re^+$,*

$$P[(x_1, \ldots, x_n) \in A] = P[(x_1, \ldots, x_n) \in A + \boldsymbol{a}]$$

for all $\boldsymbol{a} \in \Re \times \ldots \times \Re$ such that $\boldsymbol{a}^t \mathbf{1} = 0$ and $A + \boldsymbol{a}$ is an event in $\Re^+ \times \ldots \times \Re^+$, then the joint density for $x_1, \ldots, x_n$ has the form

$$p(x_1, \ldots, x_n) = \int_0^\infty \prod_{i=1}^n \theta \exp(-\theta x_i) dQ(\theta),$$

where $\theta = \lim_{n \to \infty} \bar{x}_n^{-1}$, and

$$Q(\theta) = \lim_{n \to \infty} P[(\bar{x}_n^{-1}) \leq \theta], \quad \bar{x}_n = n^{-1}(x_1 + \cdots + x_n).$$

Outline proof. (Diaconis and Ylvisaker, 1985). By the general representation theorem, there exists a random distribution function F, such that, conditional on F, $x_1, \ldots, x_n$ are independent, for any n. It can be shown that the additional invariance property continues to hold conditional on F, so that, for any $i \neq j$,

$$P[(x_i, x_j) \in A \mid F] = P[(x_i, x_j) \in A + \boldsymbol{a} \mid F]$$

for A and $\boldsymbol{a}$ as described above. If we now take $\boldsymbol{a}^t = (a_1, a_2)$ and

$$A = \{(x_i, x_j); \quad x_i > a_1 + a_2, x_j > 0\}$$

we have

$$\begin{aligned} P[(x_i > a_1 + a_2) \cap (x_j > 0) \mid F] &= P[(x_i > a_1) \cap (x_j > a_2) \mid F] \\ &= P[(x_i > a_1) \mid F] P[(x_j > a_2) \mid F]. \end{aligned}$$

By exchangeability, and recalling that x_j is certainly positive for all j, this implies that

$$P(x_i > a_1 + a_2 \mid F) = P(x_i > a_1 \mid F) P(x_i > a_2 \mid F).$$

But this functional relationship implies, for positive real-valued x_i, that

$$p(x_i > x \mid F) = e^{-\theta x}$$

for some θ, so that the density, $p(x_i \mid F) = p(x_i \mid \theta)$, is the derivative of

$$1 - \exp(-\theta x_i),$$

and hence given by $\theta \exp(-\theta x_i)$. The rest of the result follows on noting that, by the strong law of large numbers, $\theta^{-1} = \lim_{n \to \infty} [n^{-1}(x_1 + \cdots + x_n)]$. ◁

Thus, we see that judgements of exchangeability and "lack of memory" for sequences of positive real-valued random quantities constrain the possible predictive probability models for the sequence to be those which are generated by acting *as if* we have a *random sample* from an *exponential distribution* with *unknown parameter* θ, with a *prior distribution* Q for the latter. In fact, if Q^* denotes the corresponding distribution for $\phi = \theta^{-1} = \lim_{n\to\infty} \bar{x}_n$, it may be easier to use the "reparametrised" representation

$$p(x_1, \ldots, x_n) = \int_0^\infty \prod_{i=1}^n \phi^{-1} \exp(-\phi^{-1} x_i) dQ^*(\phi),$$

since Q^* is then more directly accessible as "beliefs about the sample mean from a large number of observations".

Recalling the possible motivation given above for the additional invariance assumption on the sequence $x_1, x_2, \ldots$, it is interesting to note the very specific and well-known "lack of memory" property of the exponential distribution; namely,

$$P(x_i > a_1 + a_2 \,|\, \theta, x_i > a_1) = P(x_i > a_2 \,|\, \theta),$$

which appears implicitly in the above proof.

4.4.4 The Geometric Model

Suppose $x_1, x_2, \ldots$ is judged to be an infinitely exchangeable sequence of strictly positive integer-valued random quantities. It is easy to see that we could repeat the entire introductory discussion of Section 4.4.3, except that events would now be defined in terms of sets of points on the lattice $\mathcal{Z}^+ \times \ldots \times \mathcal{Z}^+$, rather than as regions in $\Re^+ \times \ldots \times \Re^+$. This enables us to state the following representation result.

Proposition 4.8. (***Discrete representation under origin invariance***). *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of positive integer-valued random quantities with probability measure P, such that, for all n and any event A in $\mathcal{Z}^+ \times \ldots \times \mathcal{Z}^+$,*

$$P[(x_1, \ldots, x_n) \in A] = P[(x_1, \ldots, x_n) \in A + \boldsymbol{a}]$$

for all $\boldsymbol{a} \in \mathcal{Z} \times \ldots \times \mathcal{Z}$ such that $\boldsymbol{a}'\mathbf{1} = 0$ and $A + \boldsymbol{a}$ is an event in $\mathcal{Z}^+ \times \ldots \times \mathcal{Z}^+$, then the joint density for $x_1, \ldots, x_n$ has the form

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

where $\theta = \lim_{n\to\infty} \bar{x}_n^{-1}$, $\bar{x}_n = n^{-1}(x_1 + \cdots + x_n)$, and

$$Q(\theta) = \lim_{n\to\infty} P[(\bar{x}_n^{-1}) \leq \theta].$$

Outline proof. This follows precisely the steps in the proof of Proposition 4.7, except that, for positive integer-valued x_i, the functional equation

$$P(x_i > a_1 + a_2 \mid F) = P(x_i > a_1 \mid F)P(x_i > a_2 \mid F)$$

implies that

$$P(x_i > x \mid F) = \theta^x,$$

so that the probability function, $p(x_i \mid F) = p(x_i \mid \theta)$ is easily seen to be $\theta(1-\theta)^{x_i - 1}$. Again, by the strong law of large numbers, $\theta^{-1} = \lim_{n\to\infty} \bar{x}_n$, where, since $x_i \geq 1$ for all i, $0 < \theta \leq 1$. ◁

In this case, the coherent predictive probability models must be those which are generated by acting *as if* we have a *random sample* from a *geometric distribution* with *unknown parameter* θ, with a *prior distribution* Q for the latter, where $\theta^{-1} = \lim_{n\to\infty} \bar{x}_n$.

Again, recalling the possible motivation for the additional invariance property, it is interesting to note the familiar "lack of memory" property of the geometric distribution;

$$P(x_i > a_1 + a_2 \mid \theta, x_i > a_1) = P(x_i > a_2 \mid \theta).$$

4.5 MODELS VIA SUFFICIENT STATISTICS

4.5.1 Summary Statistics

We begin with a formal definition, which enables us to discuss the process of *summarising* a sequence, or *sample*, of random quantities, $x_1, \ldots, x_m$. (In general, our discussion carries over to the case of random vectors, but for notational simplicity we shall usually talk in terms of random quantities.)

Definition 4.6. (***Statistic***). *Given random quantities (vectors)* $x_1, \ldots, x_m$, *with specified sets of possible values* $X_1, \ldots, X_m$, *respectively, a random vector* $\mathbf{t}_m : X_1 \times \cdots \times X_m \to \Re^{k(m)} (k(m) \leq m)$ *is called a* $k(m)$*-dimensional statistic.*

A trivial case of such a statistic would be $\mathbf{t}_m(x_1, \ldots, x_m) = (x_1, \ldots, x_m)$, but this clearly does not achieve much by way of summarisation, since $k(m) = m$. Familiar examples of summary statistics are:

$\mathbf{t}_m = m^{-1}(x_1 + \ldots + x_m)$, *the sample mean* $(k(m) = 1)$:

$\mathbf{t}_m = [m, (x_1 + \ldots + x_m), (x_1^2 + \cdots + x_m^2)]$, *the sample size, total and sum of squares* $(k(m) = 3)$:

$\mathbf{t}_m = [m, \text{med}\{x_1, \ldots, x_m\}]$, *the sample size and median* $(k(m) = 2)$:

$\mathbf{t}_m = \max\{x_1, \ldots, x_m\} - \min\{x_1, \ldots, x_m\}$, *the sample range* $(k(m) = 1)$.

To achieve *data reduction*, we clearly need $k(m) < m$: moreover, as with the above examples, further clarity of interpretation is achieved if $k(m) = k$, a fixed dimension independent of m.

In the next section, we shall examine the formal acceptability and implications of seeking to act *as if* particular summary statistics have a special status in the context of representing beliefs about a sequence of random vectors. We shall not concern ourselves at this stage with the origin of or motivation for any such choice of particular summary statistics. Instead, we shall focus attention on the general questions of whether, and under what circumstances, it is coherent to invoke such a form of data reduction and, if so, what forms of representation for predictive probability models might result. Throughout, we shall assume that beliefs can be represented in terms of density functions.

4.5.2 Predictive Sufficiency and Parametric Sufficiency

As an example of the way in which a summary statistic might be assumed to play a special role in the evolution of beliefs, let us consider the following general situation. Past observations $x_1, \ldots, x_m$ are available and an individual is contemplating, conditional on this given information, beliefs about future observations $x_{m+1}, \ldots, x_n$, to be described by $p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m)$. The following definition describes one possible way in which assumptions of systematic data reduction might be incorporated into the structure of such conditional beliefs.

Definition 4.7. (***Predictive sufficiency***).
Given a sequence of random quantities $x_1, x_2, \ldots$, with probability measure P, where x_i takes values in $X_i, i = 1, 2, \ldots$ the sequence of statistics $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$, with $\boldsymbol{t}_j$ defined on $X_1 \times \cdots \times X_j$, is said to be predictive sufficient for the sequence $x_1, x_2, \ldots$ if, for all $m \geq 1, r \geq 1$ and $\{i_1, \ldots, i_r\} \cap \{1, \ldots, m\} = \emptyset$,

$$p(x_{i_1}, \ldots, x_{i_r} \mid x_1, \ldots, x_m) = p(x_{i_1}, \ldots, x_{i_r} \mid \boldsymbol{t}_m),$$

where $p(. \mid .)$ is the conditional density induced by P.

The above definition captures the idea that, given $\boldsymbol{t}_m = \boldsymbol{t}_m(x_1, \ldots, x_m)$, the individual values of $x_1, \ldots, x_m$ contribute nothing further to one's evaluation of probabilities of future events defined in terms of as yet unobserved random quantities. Another way of expressing this, as is easily verified from Definition 4.7, is that future observations $(x_{i_1}, \ldots, x_{i_r})$ and past observations $(x_1, \ldots, x_m)$ are conditionally independent given $\boldsymbol{t}_m$. Clearly, from a pragmatic point of view the assumption of a specified sequence of predictive sufficient statistics will, in general, greatly simplify the process of assessing probabilities of future events conditional on past observations. From a formal point of view, however, we shall need additional

structure if we are to succeed in using this idea to identify specific forms of the general representation of the joint distribution of $x_1, \ldots, x_n$.

As a *particular illustration* of what might be achieved, we shall assume in what follows that *the probability measure P describing our beliefs implies both predictive sufficiency and exchangeability for the infinite sequence* $x_1, x_2, \ldots$. As with our earlier discussion in Section 4.4, a mathematically rigorous treatment is beyond the intended level of this book and so we shall confine ourselves to an informal presentation of the main ideas.

In particular, throughout this section we shall assume that the exchangeability assumption leads to a finitely parametrised mixture representation, as in Corollary 1 to Proposition 4.3, so that, as shown in Corollary 2 to that proposition, the conditional density function of $x_{m+1}, \ldots, x_n$, given $x_1, \ldots, x_m$, has the form

$$p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m) = \int \prod_{i=m+1}^{n} p(x_i \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m)$$

where

$$dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m) = \frac{\prod_{i=1}^{m} p(x_i \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}{\int \prod_{i=1}^{m} p(x_i \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}$$

and all integrals, here and in what follows, are assumed to be over the set of possible values of $\boldsymbol{\theta}$.

This latter form makes clear that, for such exchangeable beliefs, the learning process is "transmitted" within the mixture representation by the updating of beliefs about the "unknown parameter" $\boldsymbol{\theta}$. This suggests another possible way of defining a statistic $\boldsymbol{t}_m = \boldsymbol{t}_m(x_1, \ldots, x_m)$ to be a "sufficient summary" of $x_1, \ldots, x_m$.

Definition 4.8. ***(Parametric sufficiency).*** *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of random quantities, where x_i takes values in $X_i = X, i = 1, 2, \ldots$, the sequence of statistics $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$, with $\boldsymbol{t}_j$ defined on $X_1 \times \cdots \times X_j$, is said to be parametric sufficient for $x_1, x_2, \ldots$ if, for any $n \geq 1$,*

$$dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_n) = dQ(\boldsymbol{\theta} \mid \boldsymbol{t}_n),$$

for any $dQ(\boldsymbol{\theta})$ defining an exchangeable predictive probability model via the representation

$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}) \, dQ(\boldsymbol{\theta}).$$

Definitions 4.7 and 4.8 both seem intuitively compelling as encapsulations of the notion of a statistic being a "sufficient summary". It is perhaps reassuring therefore that, within our assumed framework, we can establish the following.

Proposition 4.9. (***Equivalence of predictive and parametric sufficiencies***). *Given an infinitely exchangeable sequence of random quantities $x_1, x_2, \ldots$, where x_i takes values in $X_i = X, i = 1, 2, \ldots$, the sequence of statistics $t_1, t_2, \ldots$ with t_j defined on $X_1 \times \cdots \times X_j$ is predictive sufficient if, and only if, it is parametric sufficient.*

Heuristic proof. For any $x_1, \ldots, x_m, x_{m+1}, \ldots, x_n$ and any sequence of statistics t_m, where $t_m = t_m(x_1, \ldots, x_m), m = 1, \ldots, n-1$, the representation theorem implies that

$$p(x_{m+1}, \ldots, x_n \mid t_m) = \frac{1}{p(t_m)} p(x_{m+1}, \ldots, x_n, t_m),$$

$$= \frac{1}{p(t_m)} \int_A p(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)\, dx_1, \ldots, dx_m$$

where $A = \{(x_1 \ldots, x_m);\ t_m(x_1, \ldots, x_m) = t_m\}$, which, in turn, can be easily shown to be expressible as

$$\frac{1}{p(t_m)} \int_\Theta \left[\int_A \prod_{i=1}^{n} p(x_i \mid \theta)\, dx_1, \ldots, dx_m \right] dQ(\theta)$$

$$= \frac{1}{p(t_m)} \int_\Theta \prod_{i=m+1}^{n} p(x_i \mid \theta)\, p(t_m \mid \theta)\, dQ(\theta) = \int_\Theta \prod_{i=m+1}^{n} p(x_i \mid \theta)\, dQ(\theta \mid t_m),$$

where

$$dQ(\theta \mid t_m) = \frac{1}{p(t_m)} p(t_m \mid \theta)\, dQ(\theta) = \frac{p(t_m \mid \theta)\, dQ(\theta)}{\int p(t_m \mid \theta)\, dQ(\theta)}.$$

It follows that

$$p(x_{m+1}, \ldots, x_n \mid t_m) = p(x_{m+1}, \ldots, x_n \mid x_1, \ldots, x_m)$$

$$= \int \prod_{i=m+1}^{n} p(x_i \mid \theta)\, dQ(\theta \mid x_1, \ldots, x_m),$$

if, and only if, $dQ(\theta \mid x_1, \ldots, x_m) = dQ(\theta \mid t_m)$ for all $dQ(\theta)$. ◁

To make further progress, we now establish that parametric sufficiency is itself equivalent to certain further conditions on the probability structure.

Proposition 4.10. (***Neyman factorisation criterion***). *The sequence $t_1, t_2, \ldots$ is parametric sufficient for infinitely exchangeable $x_1, x_2, \ldots$ admitting a finitely parametrised mixture representation if and only if, for any $m \geq 1$, the joint density for $x_1, \ldots, x_m$ given θ has the form*

$$p(x_1, \ldots, x_m \mid \theta) = h_m(t_m, \theta) g(x_1, \ldots, x_m),$$

for some functions $h_m \geq 0, g > 0$.

Outline proof. Given such a factorisation, for any $dQ(\boldsymbol{\theta})$ we have

$$dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m) = \frac{p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}{\int_\Theta p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})} = \frac{h_m(\boldsymbol{t}_m, \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}{\int_\Theta h_m(\boldsymbol{t}_m, \boldsymbol{\theta}) dQ(\boldsymbol{\theta})},$$

for some $h_m > 0$. The right-hand side depends on $x_1, \ldots, x_m$ only through $\boldsymbol{t}_m$ and, hence, $dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m) = dQ(\boldsymbol{\theta} \mid \boldsymbol{t}_m)$. Conversely, given parametric sufficiency, we have, for any $dQ(\boldsymbol{\theta})$ with support Θ,

$$\frac{p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}{p(x_1, \ldots, x_m)} = dQ(\boldsymbol{\theta} \mid x_1, \ldots, x_m)$$

$$= dQ(\boldsymbol{\theta} \mid \boldsymbol{t}_m) = \frac{p(\boldsymbol{t}_m \mid \boldsymbol{\theta}) dQ(\boldsymbol{\theta})}{p(\boldsymbol{t}_m)}$$

so that

$$p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) = h_m(\boldsymbol{t}_m, \boldsymbol{\theta}) g(x_1, \ldots, x_m)$$

for some $h_m \geq 0, g > 0$ as required. ◁

Proposition 4.11. ***(Sufficiency and conditional independence).***
The sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$ is parametric sufficient for infinitely exchangeable $x_1, x_2, \ldots$ if, and only if, for any $m \geq 1$, the density $p(x_1, \ldots, x_m \mid \boldsymbol{\theta}, \boldsymbol{t}_m)$ is independent of $\boldsymbol{\theta}$.

Outline proof. For any $\boldsymbol{t}_m = \boldsymbol{t}_m(x_1, \ldots, x_m)$ we have

$$p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) = p(x_1, \ldots, x_m \mid \boldsymbol{\theta}, \boldsymbol{t}_m) p(\boldsymbol{t}_m \mid \boldsymbol{\theta}).$$

If $p(x_1, \ldots, x_m \mid \boldsymbol{\theta}, \boldsymbol{t}_m)$ is independent of $\boldsymbol{\theta}$, the parametric sufficiency of $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$ follows immediately from Proposition 4.10.

Conversely, suppose that $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$ is parametric sufficient, so that, by Proposition 4.10,

$$p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) = h_m(\boldsymbol{t}_m, \boldsymbol{\theta}) g(x_1, \ldots, x_m)$$

for some $h_m \geq 0, g > 0$. Integrating over all values $\{x_1, \ldots, x_m\}$ such that $\boldsymbol{t}_m(x_1, \ldots, x_m) = \boldsymbol{t}_m$, we obtain

$$p(\boldsymbol{t}_m \mid \boldsymbol{\theta}) = h_m(\boldsymbol{t}_m, \boldsymbol{\theta}) G(\boldsymbol{t}_m)$$

for some $G > 0$. Substituting for $h_m(\boldsymbol{t}_m, \boldsymbol{\theta})$ in the expression for $p(x_1, \ldots, x_m \mid \boldsymbol{\theta})$, we obtain

$$p(x_1, \ldots, x_m \mid \boldsymbol{\theta}) = p(\boldsymbol{t}_m \mid \boldsymbol{\theta}) \frac{g(x_1, \ldots, x_m)}{G(\boldsymbol{t}_m)}$$

so that

$$p(x_1, \ldots, x_m \mid \boldsymbol{\theta}, \boldsymbol{t}_m) = \frac{g(x_1, \ldots, x_m)}{G(\boldsymbol{t}_m)},$$

which is independent of $\boldsymbol{\theta}$. ◁

In the approach we have adopted, the definitions and consequences of predictive and parametric sufficiency have been motivated and examined within the general framework of seeking to find coherent representations of subjective beliefs about sequences of observables. Thus, for example, the notion of parametric sufficiency has so far only been put forward within the context of exchangeable beliefs, where the operational significance of "parameter" typically becomes clear from the relevant representation theorem.

In fact, however, as the reader familiar with more "conventional" approaches will have already realised, related concepts of "sufficiency" are also central to non-subjectivist theories. In particular, we note that the non-dependence of the density $p(x_1, \ldots, x_m \mid \theta, t_m)$ on θ, established here in Proposition 4.11 as a *consequence* of our definitions, was itself put forward as *the definition* of a "sufficient statistic" by Fisher (1922), and the factorisation given in Proposition 4.10 was established by Neyman (1935) as equivalent to the Fisher definition.

From an operational, subjectivist point of view, it seems to us rather mysterious to launch into fundamental definitions about learning processes expressed in terms of conditioning on "parameters" having no status other than as "labels". However, from a technical point of view, since our representation for exchangeable sequences provides, for us, a justification for regarding the usual (Fisher) definition as equivalent to predictive and parametric sufficiency, we can exploit many of the important mathematical results which have been established using that definition as a starting point.

In the context of our subjectivist discussion of beliefs and models, we shall mainly be interested in asking the following questions.

When is it coherent to act as if there is a sequence of predictive sufficient statistics associated with an exchangeable sequence of random quantities?

What forms of predictive probability model are implied in cases where we can assume a sequence of predictive sufficient statistics?

Aside from these foundational and modelling questions, however, the results given above also enable us to check the form of the predictive sufficient statistics for any given exchangeable representation. We shall illustrate this possibility with some simple examples before continuing with the general development.

Example 4.5. ***(Bernoulli model).*** We recall from Proposition 4.1 that if $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities, then we have the general representation

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_0^1 p(x_1, \ldots, x_n \mid \theta) dQ(\theta) \\ &= \int_0^1 \prod_{i=1}^{n} \mathrm{Br}(x_i \mid \theta) dQ(\theta) \\ &= \int_0^1 \theta^{s_n} (1-\theta)^{n-s_n} dQ(\theta), \end{aligned}$$

where $s_n = x_1 + \cdots + x_n$. Defining $\boldsymbol{t}_n = [n, s_n]$ and noting that we can write

$$p(x_1, \ldots, x_n \mid \theta) = h_n(\boldsymbol{t}_n, \theta) g(x_1, \ldots, x_n).$$

with

$$h_n(\boldsymbol{t}_n, \theta) = \theta^{s_n}(1-\theta)^{n-s_n}, g(x_1, \ldots, x_n) = 1,$$

it follows from Propositions 4.9 and 4.10 that the sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$ is predictive and parametric sufficient for $x_1, x_2, \ldots$. This corresponds precisely to the intuitive idea that the sequence length and total number of 1's summarises all the interesting information in any sequence of observed exchangeable 0-1 random quantities.

Example 4.6. ***(Normal model).*** We recall from Proposition 4.5 that if $x_1, x_2, \ldots$ is an exchangeable sequence of real-valued random quantities with the additional property of centred spherical symmetry then we have the general representation

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_{-\infty}^{\infty} \int_0^{\infty} p(x_1, \ldots, x_n \mid \mu, \lambda) dQ(\mu, \lambda) \\ &= \int_{-\infty}^{\infty} \int_0^{\infty} \prod_{i=1}^{n} \mathrm{N}(x_i \mid \mu, \lambda) dQ(\mu, \lambda) \\ &= \int_{-\infty}^{\infty} \int_0^{\infty} \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_i - \mu)^2\right\} dQ(\mu, \lambda) \\ &= \int_{-\infty}^{\infty} \int_0^{\infty} \left(\frac{\lambda}{2\pi}\right)^{n/2} \exp\left\{-\frac{\lambda}{2}[n(\bar{x}_n - \mu)^2 + n s_n^2]\right\} dQ(\mu, \lambda) \end{aligned}$$

where

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

In the light of Propositions 4.10 and Proposition 4.11, inspection of $p(x_1, \ldots, x_n \mid \mu, \lambda)$ reveals that

$$\boldsymbol{t}_n = (n, \bar{x}_n, s_n^2)$$

defines a sequence of predictive and parametric sufficient statistics for $x_1, x_2, \ldots$. In view of the centring and spherical symmetry conditions, it is perhaps not surprising that the sample size, mean and sample mean sum of squares about the mean turn out to be sufficient summaries. Of course, $\boldsymbol{t}_n$ is not unique; for example, since

$$n s_n^2 = x_1^2 + \cdots + x_n^2 - n(\bar{x}_n)^2$$

we could equally well define $\boldsymbol{t}_n = [n, \bar{x}_n, n^{-1}(x_1^2 + \cdots + x_n^2)]$ as the sequence of sufficient statistics.

Example 4.7. ***(Exponential model).*** We recall from Proposition 4.7 that if $x_1, x_2, \ldots$ is an exchangeable sequence of positive real-valued random quantities with an additional "origin invariance" property, then we have the general representation

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_0^\infty p(x_1, \ldots, x_n \mid \theta) dQ(\theta) \\ &= \int_0^\infty \prod_{i=1}^{n} \mathrm{Ex}(x_i \mid \theta) dQ(\theta) \\ &= \int_0^\infty \theta^n \exp(-\theta s_n) dQ(\theta) \end{aligned}$$

where $s_n = x_1 + \cdots + x_n$. Again, it is immediate from Propositions 4.10 and 4.11 that $\boldsymbol{t}_n = [n, s_n]$ defines a sequence of predictive and parametric sufficient statistics, although, in this example, there is not such an obvious link between the form of invariance assumed and the form of the sufficient statistic.

It is clear from the general definition of a sufficient statistic (parametric or predictive) that $\boldsymbol{t}_n(x_1, \ldots, x_n) = [n, (x_1, \ldots, x_n)]$ is *always* a sufficient statistic. However, given our interest in achieving simplification through data reduction, it is equally clear that we should like to focus on sufficient statistics which are, in some sense, minimal. This motivates the following definition.

Definition 4.9. ***(Minimal sufficient statistic).*** *If $x_1, x_2, \ldots$, is an infinitely exchangeable sequence of random quantities, where x_i takes values in $X_i = X$, the sequence of statistics $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$, with $\boldsymbol{t}_j$ defined on $X_1 \times \ldots \times X_j$, is minimal sufficient for $x_1, x_2, \ldots$ if given any other sequence of sufficient statistics, $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$, there exist functions $\boldsymbol{g}_1(\cdot), \boldsymbol{g}_2(\cdot), \ldots$ such that $\boldsymbol{t}_i = \boldsymbol{g}_i(\boldsymbol{s}_i)$, $i = 1, 2, \ldots$*

It is easily seen that the forms of $\boldsymbol{t}(\boldsymbol{x})$ identified in Examples 4.5 to 4.7 are minimal sufficient statistics. From now on, references to sufficient statistics should be interpreted as intending minimal sufficient statistics.

Finally, since n very often appears as part of the sufficient statistic, we shall sometimes, to avoid tedious repetition, omit explicit mention of n and refer to the "interesting function(s) of $x_1, \ldots, x_n$" as the sufficient statistic.

4.5.3 Sufficiency and the Exponential Family

In the previous section, we identified some further potential structure in the general representation of joint densities for exchangeable random quantities when predictive sufficiency is assumed. We shall now take this process a stage further by examining in detail representations relating to sufficient statistics of fixed dimension.

Since we have established, in the finite parameter framework, the equivalence of predictive and parametric sufficiency for the case of exchangeable random quantities, and their equivalence with the factorisation criterion of Proposition 4.11, we shall from now on simply use the term *sufficient statistic*, without risk of confusion.

We begin by considering exchangeable beliefs constructed by mixing, with respect to some $dQ(\theta)$, over a specified parametric form

$$p(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{k} p(x_i \mid \theta), \quad (x_1, \ldots, x_n) \in X^n \subseteq \Re^n$$

where θ is a one-dimensional parameter. By Proposition 4.10, if the form of $p(x \mid \theta)$ is such that $p(x_1, \ldots, x_n \mid \theta)$ factors into $h_n(\boldsymbol{t}_n, \theta)g(x_1, \ldots, x_n)$, for some h_n, g, the statistic $\boldsymbol{t}_n = \boldsymbol{t}_n(x_1, \ldots, x_n)$ would be sufficient. An important class of such $p(x \mid \theta)$ is identified in the following definition.

Definition 4.10. (***One-parameter exponential family***). *A probability density (or mass function) $p(x \mid \theta)$, labelled by $\theta \in \Theta \subseteq \Re$, is said to belong to the one-parameter exponential family if it is of the form*

$$p(x \mid \theta) = \mathrm{Ef}(x \mid f, g, h, \phi, \theta, c) = f(x)g(\theta)\exp\{c\phi(\theta)h(x)\}, \quad x \in X,$$

where, given f, h, ϕ, and c, $[g(\theta)]^{-1} = \int_X f(x)\exp\{c\phi(\theta)h(x)\}dx < \infty$. The family is called ***regular*** *if X does not depend on θ; otherwise it is called* ***non-regular***.

Proposition 4.12. (***Sufficient statistics for the one-parameter exponential family***). *If $x_1, x_2, \ldots, x_n \in X$, is an exchangeable sequence such that, given regular* $\mathrm{Ef}(\cdot \mid \cdot)$,

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} \mathrm{Ef}(x_i \mid f, g, h, \phi, \theta, c)dQ(\theta),$$

for some $dQ(\theta)$, then $\boldsymbol{t}_n = \boldsymbol{t}_n(x_1, \ldots, x_n) = [n, h(x_1) + \cdots + h(x_n)]$, for $n = 1, 2, \ldots$, is a sequence of sufficient statistics.

Proof. This follows immediately from Proposition 4.10 on noting that

$$\prod_{i=1}^{n} \mathrm{Ef}(x_i \mid f, g, h, \phi, \theta, c) = \prod_{i=1}^{n} f(x_i) \cdot [g(\theta)]^n \exp\left\{c\phi(\theta)\sum_{i=1}^{n} h(x_i)\right\}$$

◁

The following standard univariate probability distributions are particular cases of the (regular) one-parameter exponential family with the appropriate choices of f, g, etc. as indicated.

Bernoulli

$$p(x \mid \theta) = \mathrm{Br}(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \quad x \in \{0,1\}, \quad \theta \in [0,1].$$

$$f(x) = 1, \quad g(\theta) = 1-\theta, \quad h(x) = x, \quad \phi(\theta) = \log \frac{\theta}{1-\theta}, \quad c = 1.$$

Poisson

$$p(x \mid \theta) = \mathrm{Po}(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x \in \{0,1,2,\ldots\}, \quad \theta \in \Re^+.$$

$$f(x) = (x!)^{-1}, \quad g(\theta) = e^{-\theta}, \quad h(x) = x, \quad \phi(\theta) = \log\theta, \quad c = 1.$$

Exponential

$$p(x \mid \theta) = \mathrm{Ex}(x \mid \theta) = \theta e^{-\theta x}, \quad x \in \Re^+, \quad \theta \in \Re^+.$$

$$f(x) = 1, \quad g(\theta) = \theta, \quad h(x) = x, \quad \phi(\theta) = \theta, \quad c = -1.$$

Normal (variance unknown)

$$p(x \mid \theta) = \mathrm{N}(x \mid 0, \theta) = \left(\theta/(2\pi)\right)^{1/2} \exp\left[-\tfrac{1}{2}\theta x^2\right], \quad x \in \Re, \quad \theta \in \Re^+.$$

$$f(x) = (2\pi)^{-1/2}, \quad g(\theta) = \theta^{1/2}, \quad h(x) = x^2, \quad \phi(\theta) = \theta, \quad c = -1/2.$$

We note that the term $c\phi(\theta)$ appearing in the general $\mathrm{Ef}(\cdot \mid \cdot)$ form could always be simply written as $\phi^*(\theta)$ with ϕ^* suitably defined (see, also, Definition 4.11). However, it is often convenient to be able to separate the "interesting" function of θ, $\phi(\theta)$, from the constant which happens to multiply it.

In Definition 4.10, we allowed for the possibility (the non-regular case) that the range, X, of possible values of x might itself depend on the labelling parameter θ. Although we have not yet made a connection between this case and forms of representation arising in the modelling of exchangeable sequences, it will be useful at this stage to note examples of the well-known forms of distribution which are covered by this definition. We shall indicate later how the use of such forms in the modelling process might be given a subjectivist justification.

Uniform

$$p(x \mid \theta) = \mathrm{U}(x \mid 0, \theta) = \theta^{-1}, \quad x \in (0, \theta), \quad \theta \in \Re^+.$$

$$f(x) = 1, \quad g(\theta) = \theta^{-1}, \quad h(x) = 0, \quad \phi(\theta) = \theta, \quad c = 1.$$

Shifted exponential

$$p(x \mid \theta) = \text{Shex}(x \mid \theta) = \exp[-(x-\theta)], \quad x - \theta \in \Re^+, \quad \theta \in \Re^-.$$

$$f(x) = e^{-x}, \quad g(\theta) = e^{\theta}, \quad h(x) = 0, \quad \phi(\theta) = \theta, \quad c = 1.$$

In order to identify sequences of sufficient statistics in these and similar cases, we make use of the factorisation criterion given in Proposition 4.10.

For the uniform, we rewrite the density in the form

$$p(x \mid \theta) = \theta^{-1} I_{(0,\theta)}(x), \qquad x \in \Re,$$

so that, for any sequence $x_1, \ldots, x_n$ which is conditionally independent given θ,

$$\begin{aligned} p(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} p(x_i \mid \theta) \\ &= \theta^{-n} I_{(0,\theta)}\Big(\max_{i=1,\ldots,n} \{x_i\}\Big), \qquad (x_1, \ldots, x_n) \in \Re^n. \end{aligned}$$

It then follows immediately from Proposition 4.10 that

$$t_n = t_n(x_1, \ldots, x_n) = \left[n, \max_{i=1,\ldots,n} \{x_i\}\right], \qquad n = 1, 2, \ldots$$

is a sequence of sufficient statistics in this case.

For the shifted exponential, if we rewrite the density in the form

$$p(x \mid \theta) = \exp[-(x-\theta)]\, I_{(\theta,\infty)}(x), \qquad x \in \Re.$$

a similar argument shows that, for $(x_1, \ldots, x_n) \in \Re^+$,

$$p(x_1, \ldots, x_n \mid \theta) = \exp[-n\bar{x}_n] \exp[n\theta]\, I_{(\theta,\infty)}\Big(\min_{i=1,\ldots,n} \{x_i\}\Big),$$

so that, for $n = 1, 2, \ldots$,

$$t_n = t_n(x_1, \ldots, x_n) = \left[n, \min_{i=1,\ldots,n} \{x_i\}\right]$$

provides a sequence of sufficient statistics.

The above discussion readily generalises to the case of exchangeable sequences generated by mixing over specified parametric forms involving a k-dimensional parameter $\boldsymbol{\theta}$.

Definition 4.11. (***k*-parameter exponential family**). *A probability density (or mass function) $p(x \mid \boldsymbol{\theta})$, $x \in X$, which is labelled by $\boldsymbol{\theta} \in \Theta \subseteq \Re^k$, is said to belong to the k-parameter exponential family if it is of the form*

$$p(x \mid \boldsymbol{\theta}) = \mathrm{Ef}_k(x \mid f, g, \boldsymbol{h}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}) = f(x)g(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) h_i(x)\right\},$$

where $\boldsymbol{h} = (h_1, \ldots, h_k)$, $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\phi_1, \ldots, \phi_k)$ and, given the functions $f, \boldsymbol{h}, \boldsymbol{\phi}$, and the constants c_i,

$$\frac{1}{g(\boldsymbol{\theta})} = \int_X f(x) \exp\left\{\sum_{i=1}^{k} c_i \phi_i(\boldsymbol{\theta}) h_i(x)\right\} dx < \infty.$$

*The family is called **regular** if X does not depend on $\boldsymbol{\theta}$; otherwise it is called **non-regular**.*

Proposition 4.13. (***Sufficient statistics for the k-parameter exponential family**). If $x_1, x_2, \ldots, x_i \in X$, is an exchangeable sequence such that, given regular k-parameter $\mathrm{Ef}_k(\cdot \mid \cdot)$,*

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} \mathrm{Ef}_k(x_i \mid f, g, \boldsymbol{h}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}) dQ(\boldsymbol{\theta}),$$

for some $dQ(\boldsymbol{\theta})$, then

$$\boldsymbol{t}_n = \boldsymbol{t}_n(x_1, \ldots, x_n) = \left[n, \sum_{i=1}^{n} h_1(x_i), \ldots, \sum_{i=1}^{n} h_k(x_i)\right], n = 1, 2, \ldots$$

is a sequence of sufficient statistics.

Proof. This is analogous to Proposition 4.12 and is a straightforward consequence of Proposition 4.10. ◁

The following standard probability distributions are particular cases (the first regular, the second non-regular) of the k-parameter exponential family with the appropriate choices of f, g etc. as indicated.

Normal (unknown mean and variance)

$$p(x \mid \boldsymbol{\theta}) = p(x \mid \mu, \tau) = \mathrm{N}(x \mid \mu, \tau)$$
$$= \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{\tau}{2}(x-\mu)^2\right], \quad x \in \Re, \quad \mu \in \Re, \quad \tau \in \Re^+.$$

In this case, $k = 2$ and

$$f(x) = (2\pi)^{-1/2}, \quad g(\boldsymbol{\theta}) = \tau^{1/2} \exp[-\tfrac{1}{2}\tau\mu^2], \quad \boldsymbol{h}(x) = (x, x^2),$$

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = (\tau\mu, \tau), \quad c_1 = 1, \quad c_2 = -1/2,$$

so that $\boldsymbol{t}_n = \left[n, \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right], n = 1, 2, \ldots$ is a sequence of sufficient statistics.

Uniform (over the interval $[\theta_1, \theta_2]$)

$$p(x \mid \boldsymbol{\theta}) = p(x \mid \theta_1, \theta_2) = \mathrm{U}(x \mid \theta_1, \theta_2) = (\theta_2 - \theta_1)^{-1},$$
$$x \in (\theta_1, \theta_2), \quad \theta_1 \in \Re, \quad \theta_2 - \theta_1 \in \Re^+.$$

In this case,

$$f(x) = 1, \quad g(\boldsymbol{\theta}) = (\theta_2 - \theta_1)^{-1}, \quad \boldsymbol{h}(x) = 0, \quad \boldsymbol{\phi}(\boldsymbol{\theta}) = (\theta_1, \theta_2), \quad c_1 = c_2 = 0,$$

and

$$\boldsymbol{t}_n = [n, \min\{x_1, \ldots, x_n\}, \max\{x_1, \ldots, x_n\}], n = 1, 2, \ldots$$

is easily seen to give a sequence of sufficient statistics.

The description of the exponential family forms given in Definitions 4.10 and 4.11, is convenient for some purposes (relating straightforwardly to familiar versions of parametric families), but somewhat cumbersome for others. This motivates the following definition, which we give for the general k-parameter case.

Definition 4.12. ***(Canonical exponential family).***
The probability density (or mass function)

$$p(\boldsymbol{y} \mid \boldsymbol{\psi}) = \mathrm{Cef}(\boldsymbol{y} \mid a, b, \boldsymbol{\psi}) = a(\boldsymbol{y}) \exp\{\boldsymbol{y}'\boldsymbol{\psi} - b(\boldsymbol{\psi})\}, \qquad \boldsymbol{y} \in Y,$$

derived from $\mathrm{Ef}_k(\cdot \mid \cdot)$ *in Definition 4.11, via the transformations*

$$\boldsymbol{y} = (y_1, \ldots, y_k), \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_k),$$

$$y_i = h_i(x), \qquad \psi_i = c_i \phi_i(\boldsymbol{\theta}), \qquad i = 1, \ldots, k,$$

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

Systematic use of this canonical form to clarify the nature of the Bayesian learning process will be presented in Section 5.2.2. Here, we shall use it to examine briefly the nature and interpretation of the function $b(\psi)$, and to identify the distribution of sums of independent Cef random quantities.

Proposition 4.14. (***First two moments of the canonical exponential family***). *For y in Definition 4.12,*

$$E(y \mid \psi) = \nabla b(\psi), \qquad V(y \mid \psi) = \nabla^2 b(\psi).$$

Proof. It is easy to verify that the characteristic function of y conditional on ψ is given by

$$E(\exp\{iu^t y\} \mid \psi) = \exp\{b(iu + \psi) - b(\psi)\},$$

from which the result follows straightforwardly. ◁

Proposition 4.15. (***Sufficiency in the canonical exponential family***). *If $y_1, \cdots, y_n$ are independent* $\mathrm{Cef}(y \mid a, b, \psi)$ *random quantities, then*

$$s = \sum_{i=1}^{n} y_i$$

is a sufficient statistic and has a distribution $\mathrm{Cef}(s \mid a^{(n)}, nb, \psi)$, *where $a^{(n)}$ is the n-fold convolution of a.*

Proof. Sufficiency is immediate from Proposition 4.12. We see immediately that the characteristic function of s is $\exp\{nb(iu + \psi) - nb(\psi)\}$, so that the distribution of s is as claimed, where $a^{(n)}$ satisfies

$$nb(\psi) = \log \int a^{(n)}(s) \exp\{\psi^t s\} ds.$$

Examination of the density convolution form for $n = 1$, plus induction, establishes the form of $a^{(n)}$. ◁

Our discussion thus far has considered the situation where exchangeable belief distributions are constructed by assuming a mixing over finite-parameter exponential family forms. A consequence is that sufficient statistics of fixed dimension exist. Moreover, classical results of Darmois (1936), Koopman (1936), Pitman (1936), Hipp (1974) and Huzurbazar (1976) establish, under various regularity conditions, that the exponential family is the only family of distributions for which such sufficient statistics exist.

In the second part of this subsection, we shall consider the question of whether there are structural assumptions about an exchangeable sequence $x_1, x_2, \ldots$ which *imply* that the mixing *must* be over exponential family forms.

Previously, in Section 4.4, we considered particular *invariance* assumptions, which, together with exchangeability, identified the parametric forms that had to appear in the mixture representation. Here, we shall consider, instead, whether characterisations can be established via assumptions about *conditional distributions*, motivated by *sufficiency* ideas.

As a preliminary, suppose for a moment that an exchangeable sequence, $\{\boldsymbol{y}_i\}$, is modelled by

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \int \prod_{i=1}^{n} \mathrm{Cef}(\boldsymbol{y}_i \mid a, b, \boldsymbol{\psi}) dQ(\boldsymbol{\psi}).$$

Now consider the form of $p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n = \boldsymbol{s})$, $k < n$. Because of exchangeability, this has a representation as a mixture over

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n = \boldsymbol{s}, \boldsymbol{\psi}).$$

But the latter does not involve $\boldsymbol{\psi}$ because of the sufficiency of $\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n$ (Propositions 4.11 and 4.15), so that

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \textstyle\sum_{i=1}^{n} \boldsymbol{y}_i = \boldsymbol{s}) = p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k \mid \textstyle\sum_{i=1}^{n} \boldsymbol{y}_i = \boldsymbol{s}, \boldsymbol{\psi})$$

$$= \prod_{i=1}^{k} a(\boldsymbol{y}_i) \exp\{\boldsymbol{\psi}^t \boldsymbol{y}_i - b(\boldsymbol{\psi})\} \frac{a^{(n-k)}(\boldsymbol{s} - \boldsymbol{s}_k) \exp\{\boldsymbol{\psi}^t(\boldsymbol{s} - \boldsymbol{s}_k) - (n-k)b(\boldsymbol{\psi})\}}{a^{(n)}(\boldsymbol{s}) \exp\{\boldsymbol{\psi}^t \boldsymbol{s} - nb(\boldsymbol{\psi})\}}$$

where, in the numerator, $\boldsymbol{s}_k = \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_k \leq \boldsymbol{s}$. The exponential family mixture representation thus *implies* that,

$$p(\textstyle\sum_{i=1}^{k} \boldsymbol{y}_i \mid \sum_{i=1}^{n} \boldsymbol{y}_i = \boldsymbol{s}) = \dfrac{\prod_{i=1}^{k} a(\boldsymbol{y}_i) a^{(n-k)}(\boldsymbol{s} - \boldsymbol{s}_k)}{a^{(n)}(\boldsymbol{s})}.$$

Now suppose we consider the converse. If we assume $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ to be exchangeable and also assume that, for all n and $k < n$, the conditional distributions have the above form (for some a defining a $\mathrm{Cef}(\boldsymbol{y} \mid a, b, \boldsymbol{\psi})$ form), does this imply that $p(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ has the corresponding exponential family mixture form? A rigorous mathematical discussion of this question is beyond the scope of this volume (see Diaconis and Freedman, 1990). However, with considerable licence in ignoring regularity conditions, the result and the "flavour" of a proof are given by the following.

Proposition 4.16. (***Representation theorem under sufficiency***). *If $y_1, y_2, \ldots$ is any exchangeable sequence such that, for all $n \geq 2$ and $k < n$,*

$$p(y_1, \ldots, y_k \mid y_1 + \cdots + y_n = s) = \prod_{i=1}^{k} a(y_i) a^{(n-k)}(s - s_k)/a^{(n)}(s),$$

where $s_k = y_1 + \cdots + y_k$ and $a(\cdot)$ defines $\mathrm{Cef}(y \mid a, b, \psi)$, then

$$p(y_1, \ldots, y_n) = \int \prod_{i=1}^{n} \mathrm{Cef}(y_i \mid a, b, \psi) dQ(\psi),$$

for some $dQ(\psi)$.

Outline proof. We first note that exchangeability implies a mixture representation, mixing over distributions which make the y_i independent. But each of the latter distributions, with densities denoted generically by f, themselves imply an exchangeable sequence, so that, for $n \geq 2, k < n$, $f(y_1, \ldots, y_k \mid y_1 + \cdots + y_n = s)$ also has the specified form in terms of $a(\cdot)$.

Now consider $n = 2, k = 1$. Independence implies that

$$f(y_1 \mid y_1 + y_2 = s) = \frac{f(y_1) f(s - y_1)}{f^{(2)}(s)},$$

where $f(\cdot)$ denotes the marginal density and $f^{(2)}(\cdot)$ its twofold convolution, so that $f(\cdot)$ must satisfy

$$\frac{f(y_1) f(s - y_1)}{f^{(2)}(s)} = \frac{a(y_1) a(s - y_1)}{a^{(2)}(s)}.$$

If we now define

$$u(y_1) = \log \frac{f(y_1)}{a(y_1)} - \log \frac{f(0)}{a(0)}$$

and

$$v(s) = \log \frac{f^{(2)}(s)}{a^{(2)}(s)} - 2 \log \frac{f(0)}{a(0)},$$

it follows that

$$u(y_1) + u(s - y_1) = v(s).$$

Setting $y_1 = s$, and noting that $u(0) = 0$, we obtain $u(s) = v(s)$, and hence

$$u(y_1) + u(y_2) = u(y_1 + y_2).$$

This implies that $u(y) = \psi' y$, for some ψ, so that

$$f(y) = a(y) \exp\{\psi' y - b(\psi)\}.$$ ◁

The following example provides a concrete illustration of the general result.

Example 4.8. ***(Characterisation of the Poisson model).*** Suppose that the sequence of non-negative integer valued random quantities $y_1, y_2, \ldots$ is judged exchangeable, with the conditional distribution of $\boldsymbol{y} = (y_1, \ldots, y_k)$ given $y_1 + \cdots + y_n = s$, $n \geq 2, k < n$, specified to be the multinomial $\mathrm{Mu}_k(\boldsymbol{y} \mid s, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (1/n, \ldots, 1/n)$, so that

$$p(y_1, \ldots, y_k \mid y_1 + \cdots + y_n = s) = \frac{s!}{\prod_{i=1}^{k} y_i!(s - s_k)!} \prod_{i=1}^{k} \left(\frac{1}{n}\right)^{y_i} \left(1 - \frac{k}{n}\right)^{s - s_k},$$

where $s_k = y_1 + \cdots + y_k$. Noting that the Poisson distribution, $\mathrm{Pn}(y \mid v)$, can be written in $\mathrm{Cef}(y \mid a, b, v)$ form as

$$\mathrm{Pn}(y \mid v) = \frac{1}{y!} \exp\{yv - e^v\} = a(y) \exp\{yv - b(v)\},$$

from which it easily follows that $a^{(n)}(s) = n^s/s!$, it is straightforward to check that, in terms of $a(\cdot)$ and $a^{(n)}(\cdot)$,

$$M_k(\boldsymbol{y} \mid s, \boldsymbol{\theta}) = \prod_{i=1}^{k} a(y_i) a^{(n-k)}(s - s_k)/a^{(n)}(s).$$

By Proposition 4.16, it follows that the belief specification for $y_1, y_2, \ldots$ is coherent and implies that

$$p(y_1, \ldots, y_n) = \int_0^\infty \prod_{i=1}^{n} \mathrm{Pn}(y_i \mid v) dQ(v),$$

for some $dQ(v)$, $v \in \Re^+$. ◁

As we remarked earlier, the above heuristic analysis and discussion for the k-parameter regular exponential family has been given without any attempt at rigour. For the full story the reader is referred to Diaconis and Freedman (1990). Other relevant references for the mathematics of exponential families include Barndorff-Nielsen (1978), Morris (1982) and Brown (1985).

We conclude this subsection by considering, briefly and informally, what can be said about characterisations of exchangeable sequences as mixtures of non-regular exponential families. For concreteness, we shall focus on the uniform, $U(x \mid 0, \theta)$, distribution, which has density $\theta^{-1} I_{(0,\theta)}(x)$, $x \in \Re$, and sufficient statistic $\max\{x_1, \ldots, x_n\}$, given a sample $x_1, \ldots, x_n$. This sufficient statistic is clearly not a summation, as is the case for regular families (and plays a key role in Proposition 4.16). However, conditional on $m_n = \max\{x_1, \ldots, x_n\}$, $x_1, \ldots, x_k$, $k << n$, are approximately independent $U(x_i \mid 0, m_n)$ and this will therefore be true for all exchangeable $x_1, x_2, \ldots$ constructed by mixing over independent $U(x_i \mid 0, \theta)$. Conversely, we might wonder whether positive exchangeable sequences having

this conditional property are necessarily mixtures of independent $U(x_i \mid 0, \theta)$. Intuitively, if m_n tends to a finite θ from below, as $n \to \infty$, one might expect the result to be true. This is indeed the case, but a general account of the required mathematical results is beyond our intended scope in this volume. The interested reader is referred to Diaconis and Freedman (1984), and the further references discussed in Section 4.8.1.

4.5.4 Information Measures and the Exponential Family

Our approach to the exponential family has been through the concept of predictive or, equivalently, parametric sufficient statistics. It is interesting to note, however, that exponential family distributions can also be motivated through the concept of the utility of a distribution (c.f. Section 3.4), using the derived notions of approximation and discrepancy.

Consider the following problem. We seek to obtain a mathematical representation of a probability density $p(x)$, which satisfies the k (independent) constraints

$$\int_X h_i(x)p(x)dx = m_i < \infty, \quad i = 1, \ldots, k,$$

where $m_1, \ldots, m_k$ are specified constants, together with the normalizing constraint $\int_X p(x)dx = 1$, and, in addition, is to be approximated as closely as possible by a specified density $f(x)$.

We recall from Definition 3.20 (with a convenient change of notation) that the *discrepancy* from a probability density $p(x)$ *assumed to be true* of an *approximation* $f(x)$ is given by

$$\delta(f \mid p) = \int_X p(x) \log \frac{p(x)}{f(x)} dx,$$

where f and p are both assumed to be strictly positive densities over the same range, X, of possible values. Note that we are interested in deriving a mathematical representation of the *true* probability density $p(x)$, not of the (specified) approximation $f(x)$. Thus, we minimise $\delta(f \mid p)$ over p subject to the required constraints on p, rather than $\delta(f \mid p)$ over f subject to constraints on f. Hence, we seek p to minimise

$$F(p) = \int_X p(x) \log \frac{p(x)}{f(x)} dx + \sum_{i=1}^{k} \theta_i \left[\int_X h_i(x)p(x)dx - m_i \right] + c \left[\int_X p(x)dx - 1 \right],$$

where $\theta_1, \ldots, \theta_k$ and c are arbitrary constant multipliers.

Proposition 4.17. (***The exponential family as an approximation***). *The functional $F(p)$ defined above is minimised by*

$$p(x) = \mathrm{Ef}_k(x \mid f, g, \boldsymbol{h}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}), x \in X$$

where f and $\boldsymbol{h}$ are given in $F(p)$, $c_i = 1$, $\boldsymbol{\phi} = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ and

$$\frac{1}{g(\boldsymbol{\theta})} = \int_X f(x) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(x)\right\} dx.$$

Proof. By a standard variational argument (see, for example, Jeffreys and Jeffreys, 1946, Chapter 10), a necessary condition for p to give a stationary value of $F(p)$ is that

$$(\partial/\partial\alpha) F(p(x) + \alpha\tau(x)) \mid_{\alpha=0} = 0$$

for any function $\tau : x \to \Re$ of sufficiently small norm. This condition reduces to the equation

$$\int \left[\log(p(x)/f(x)) + \sum_{i=1}^{k} \theta_i h_i(x) + (c+1)\right] \tau(x) dx = 0,$$

from which it follows that

$$p(x) \propto f(x) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(x)\right\},$$

as required. (For an alternative proof, see Kullback, 1959/1968, Chapter 3.) ◁

The resulting exponential family form for $p(x)$ was derived on the basis of a given approximation $f(x)$ and a collection of "constant" functions $\boldsymbol{h}(x) = [h_i(x), \ldots, h_k(x)]$. If we wish to emphasise this derivation of the family, we shall refer to $\mathrm{Ef}(x \mid f, g, \boldsymbol{h}, \phi, \theta, \boldsymbol{c})$ as *the exponential family generated by f and $\boldsymbol{h}$.*

In general, specification of the sufficient statistic

$$\boldsymbol{t}_m = \left[m, \sum_{i=1}^{m} h_i(x_i), \ldots, \sum_{i=1}^{m} h_k(x_i)\right]$$

does not uniquely identify the form of $f(x)$ within the exponential family framework. Consider, for example, the $\mathrm{Ga}(x \mid \alpha, \theta)$ family with α known. Each distinct α defines a distinct exponential family with density

$$(x^{\alpha-1}/\Gamma(\alpha))\theta^{\alpha} \exp\{-\theta x\}.$$

so that, in addition to $h(x) = x$, we need to specify $f(x) = x^{\alpha-1}/\Gamma(\alpha)$ in order to identify the family.

Returning to the general problem of choosing p to be "as close as possible" to an "approximation" f, subject to the k constraints defined by $\boldsymbol{h}(x)$, it is interesting to ask what happens if the approximation f is very "vague", in the sense that f is extremely diffusely spread over X. A limiting form of this would be to consider $f(x) = \text{constant}$, which leads us to seek the p minimising $\int_X p(x) \log p(x) dx$ subject to the given constraints. The solution is then

$$p(x) = \frac{\exp\left\{\sum_{i=1}^{k} \theta_i h_i(x)\right\}}{\int_X \exp\left\{\sum_{i=1}^{k} \theta_i h_i(x)\right\} dx}$$

which, since minimising $\int_X p(x) \log p(x)\, dx$ is equivalent to maximising $H(p) = -\int_X p(x) \log p(x)\, dx$, is the so-called *maximum entropy* choice of p.

Thus, for example, if $X = \Re^+$ and $h(x) = x$, the maximum entropy choice for $p(x)$ is $\text{Ex}(x \mid \phi)$, the exponential distribution with $\phi^{-1} = \text{E}(x \mid \phi)$. If $X = \Re$ and $\boldsymbol{h}(x) = (x, x^2)$, the maximum entropy choice for $p(x)$ turns out to be $\text{N}(x \mid \mu, \lambda)$, the normal distribution with $\mu = E(x \mid \mu, \lambda)$, $\lambda^{-1} = V(x \mid \mu, \lambda)$ (c.f. Example 3.4, following Definition 3.20).

Our discussion of modelling has so far concentrated on the case of beliefs about a single sequence of observations $x_1, x_2, \ldots$, judged to have various kinds of invariance or sufficiency properties. In the next section, we shall extend our discussion in order to relate these ideas to the more complex situations, which arise when several such sequences of observations are involved, or when there are several possible ways of making exchangeable or related judgements about sequences.

4.6 MODELS VIA PARTIAL EXCHANGEABILITY

4.6.1 Models for Extended Data Structures

In Section 4.5, we discussed various kinds of justification for modelling a sequence of random quantities $x_1, x_2, \ldots$ as a random sample from a parametric family with density $p(x \mid \theta)$, together with a prior distribution $dQ(\theta)$ for θ. We also briefly mentioned further possible kinds of judgements, involving assumptions about conditional moments or information considerations, which further help to pinpoint the appropriate specification of a parametric family.

However, in order to concentrate on the basic conceptual issues, we have thus far restricted attention to the case of a *single sequence* of random quantities, $x_1, x_2, \ldots$, labelled by a *single index*, $i = 1, 2, \ldots$, and *unrelated to other random quantities*. Clearly, in many (if not most) areas of application of statistical modelling the situation will be more complicated than this, and we shall need to extend and

adapt the basic form of representation to deal with the perceived complexities of the situation. Among the typical (but by no means exhaustive) kinds of situation we shall wish to consider are the following.

(i) Sequences $x_{i1}, x_{i2}, \ldots$ of random quantities are to be observed in each of $i \in I$ contexts. For example: we may have sequences of clinical responses to each of I different drugs; or responses to the same drug used on I different subgroups of a population. A modelling framework is required which enables us to learn, in some sense, about differences between some aspect of the responses in the different sequences.

(ii) In each of $i \in I$ contexts, $j \in J$ different treatments are each replicated $k \in K$ times, and the random quantities x_{ijk} denote observable responses for each context/treatment/replicate combination. For example: we may have I different irrigation systems for fruit trees, J different tree pruning regimes and K trees exposed to each irrigation/pruning combination, with x_{ijk} denoting the total yield of fruit in a given year; or we may have I different geographical areas, J different age-groups and K individuals in each of the IJ combinations, with x_{ijk} denoting the presence or absence of a specific type of disease, or a coding of voting intention, or whatever. A modelling framework is required which enables us to investigate differences between either contexts, or treatments, or context/treatment combinations.

(iii) Sequences of random quantities $x_{i1}, x_{i2}, \ldots, i \in I$, are to be observed, where some form of qualitative assumption has been made about a form of relationship between the x_{ij} and other specified (controlled or observed) quantities $z_i = (z_{i1}, \ldots, z_{ik}), k \geq 1$. For example: x_{ij} might denote the status (dead or alive) of the jth rat exposed to a toxic substance administered at dose level z_i, with an assumed form of relationship between z_i and the corresponding "death rate"; or x_{ij} might denote the height or weight at time z_i from the jth replicate measurement of a plant or animal following some assumed form of "growth curve"; or x_{ij} might denote the output yield on the jth run of a chemical process when k inputs are set at the levels $z_i = (z_{i1}, \ldots, z_{ik})$ and the general form of relationship between process output and inputs is either assumed known or well-approximated by a specified mathematical form. In each case a modelling framework is required which enables us to learn about the quantitative form of the relationship, and to quantify beliefs (predictions) about the observable x^* corresponding to a specified input or control quantity z^*.

(iv) Exchangeable sequences, $x_{i1}, x_{i2}, \ldots$, of random quantities are to be observed in each of $i \in I$ contexts, where I is itself a selection from a potentially larger index set I^*. Suppose that for each sequence,

$$t_m^{(i)} = \left[m, \sum_{j=1}^{m} s^{(i)}(x_j) \right], \quad i \in I,$$

is judged to be a sufficient statistic, that the strong law limits

$$\theta_i = \lim_{i \to \infty} \frac{1}{m} t_m^{(i)}, \quad i \in I,$$

exist and that the sequence $\theta_1, \theta_2, \ldots$ is itself judged exchangeable. For example: sequence i may consist of 0 – 1 (success-failure) outcomes on repeated trials with the ith of I similar electronic components; or sequence i may consist of quality measurements of known precision on replicate samples of the ith of I chemically similar dyestuffs. In the first case, the sequence of long-run frequencies of failures for each of the components might, a priori, be judged to be exchangeable; in the second case, the sequence of large-sample averages of quality for each of the dyestuffs might, a priori, be judged to be exchangeable. A modelling framework is required which enables us to exploit such further judgements of exchangeability in order to be able to use information from *all* the sequences to strengthen, in some sense, the learning process within an *individual* sequence.

4.6.2 Several Samples

We shall begin our discussion of possible forms of partial exchangeability judgements for several sequences of observables, $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$, by considering the simple case of 0 – 1 random quantities.

In many situations, including that of a comparative clinical trial, joint beliefs about several sequences of 0 – 1 observables would typically have the property encapsulated in the following definition, where, here and throughout this section, $\boldsymbol{x}_i(n_i)$ denotes the vector of random quantities $(x_{i1}, \ldots, x_{in_i})$.

Definition 4.13. (***Unrestricted exchangeability for 0 – 1 sequences***). *Sequences of 0 – 1 random quantities, $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$, are said to be unrestrictedly exchangeable if each sequence is infinitely exchangeable and, in addition, for all $n_i \leq N_i, i = 1, \ldots, m$,*

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m) \mid y_1(N_1), \ldots, y_m(N_m)) = \prod_{i=1}^{m} p(\boldsymbol{x}_i(n_i) \mid y_i(N_i)),$$

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

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, $i = 1, \ldots m$, *only the total for the ith 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. Thus, 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 ($N_2 = 80$, $y_2(N_2) = 20$), we would typically judge the latter information to be irrelevant to any assessment of the probability that the first three patients receiving Drug 1 survived and the fourth one died ($x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{14} = 1$). Of course, the information might well be judged relevant if we were *not* informed of the value of $y_1(N_1)$. The definition thus encapsulates a kind of "conditional irrelevance" judgement.

As an example of a situation where this condition does *not* apply, suppose that $x_{11}, x_{12}, \ldots$ is an infinitely exchangeable 0–1 sequence and that another sequence $x_{21}, x_{22}, \ldots$ is defined by $x_{2j} = x_{1j}$ (or by $x_{2j} = 1 - x_{1j}$). Then $x_{21}, x_{22}, \ldots$ is certainly an exchangeable sequence (since $x_{11}, x_{12}, \ldots$ is), but, taking $x_{2j} = x_{1j}$ and $n_1 = n_2 = N_1 = N_2 = 2$,

$$p(x_{11} = 0, x_{12} = 1, x_{21} = 1, x_{22} = 0 \mid y_{12} = 1, y_{22} = 1) = 0.$$

whereas

$$p(x_{11} = 0, x_{12} = 1 \mid y_{12} = 1)\, p(x_{21} = 1, x_{22} = 0 \mid y_{22} = 1) = 1/2 \times 1/2 = 1/4.$$

Further insight is obtained by noting (from Definition 4.13) that unrestricted exchangeability implies that

$$p(x_{11}, \ldots, x_{1n_1}, \ldots, x_{m1}, \ldots, x_{mn_m}) = p(x_{1\pi_1(1)}, \ldots, x_{1\pi_1(n_1)}, \ldots, x_{m\pi_m(1)}, \ldots, x_{m\pi_m(n_m)})$$

for any *unrestricted* choice of permutations π_i of $\{1, \ldots, n_i\}$, $i = 1, \ldots, m$, whereas, in the case of the above counter-example, we only have invariance of the joint distribution when $\pi_1 = \pi_2$. For a development starting from this latter condition see de Finetti (1938).

We can now establish the following generalisation of Proposition 4.1.

Proposition 4.18. (***Representation theorem for several sequences of 0–1 random quantities***). *If $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$ are unrestrictedly infinitely exchangeable sequences of 0–1 random quantities with joint probability measure P, there exists a distribution function Q such that*

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

where, with $y_i(n_i) = x_{i1} + \cdots + x_{in_i}$, $i = 1, \ldots, m$,

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

Corollary. *Under the conditions of Proposition 4.18,*

$$p(y_1(n_1),\ldots,y_m(n_m)) = \int_{[0,1]^m} \prod_{i=1}^{m} \binom{n_i}{y_i(n_i)} \theta_i^{y_i(n_i)}(1-\theta_i)^{n_i-y_i(n_i)} dQ(\theta_1,\ldots,\theta_m)$$

Proof. We first note that

$$p(y_1(n_1),\ldots,y_m(n_m)) = \binom{n_1}{y_1(n_1)} \cdots \binom{n_m}{y_m(n_m)} p(\boldsymbol{x}_1(n_1),\ldots,\boldsymbol{x}_m(n_m))$$

so that, to prove the proposition, it suffices to establish the corollary. Moreover, for any $N_i \geq n_i$, $i = 1,\ldots,m$, we may express $p(y_1(n_1),\ldots,y_m(n_m))$ as

$$\sum p(y_1(n_1),\ldots,y_m(n_m) \mid y_1(N_1),\ldots,y_m(N_m))\; p(y_1(N_1),\ldots,y_m(N_m)),$$

where the ith of the m summations ranges from $y_i(N_i) = y_i(n_i)$ to $y_i(N_i) = N_i$, and where, by Definition 4.4 and a straightforward generalisation of the argument given in Proposition 4.1,

$$p(y_1(n_1),\ldots,y_m(n_m) \mid y_1(N_1),\ldots,y_m(N_m)) = \prod_{i=1}^{m} p(y_i(n_i) \mid y_i(N_i)) = \prod_{i=1}^{m} \binom{n_i}{y_i(n_i)} \binom{N_i - n_i}{n_i - y_i(n_i)} \Big/ \binom{N_i}{n_i}.$$

Writing $(y_N)_{y_n} = y_N(y_N-1)\cdots(y_N-(y_n-1))$, etc., and defining the function $Q_{N_1,\ldots,N_m}(\theta_1,\ldots,\theta_m)$ on $\Re^m$ to be the m-dimensional "step" function with "jumps" of $p(y_1(N_1),\ldots,y_m(N_m))$ at

$$(\theta_1,\ldots,\theta_m) = \left(\frac{y_1(N_1)}{N_1},\ldots,\frac{y_m(N_m)}{N_m}\right),$$

where $y_i(N_i) = 0,\ldots,N_i$, $i = 1,\ldots,m$. We see that $p(y_1(n_1),\ldots,y_m(n_m))$ is equal to

$$\int_{[0,1]^m} \prod_{i=1}^{m} \left\{ \binom{n_i}{y_i(n_i)} \frac{(\theta_i N_i)_{y_i(n_i)}[(1-\theta_i)N_i]_{n_i-y_i(n_i)}}{(N_i)_{n_i}} \right\} dQ_{N_1,\ldots,N_m}(\boldsymbol{\theta}).$$

As $N_1,\ldots,N_m \to \infty$,

$$\prod_{i=1}^{m} \left\{ \frac{(\theta_i N_i)_{y_i(n_i)}[(1-\theta_i)N_i]_{n_i-y_i(n_i)}}{(N_i)_{n_i}} \right\} \to \prod_{i=1}^{m} \theta_i^{y_i(n_i)}\,(1-\theta_i)^{n_i-y_i(n_i)},$$

uniformly in $\theta_1,\ldots,\theta_m$, and, by the multidimensional version of Helly's theorem (see Section 3.2.3), there exists a subsequence $Q_{N_1(j),\ldots,N_m(j)}$, $j = 1,2,\ldots$ having a limit Q, which is a distribution function on $\Re^m$. The result follows. ◁

Considering, for simplicity, $m = 2$, Proposition 4.18 (or its corollary) asserts that if we judge two sequences of 0 – 1 random quantities to be unrestrictedly exchangeable, we can proceed *as if*:

(i) the x_{ij} are judged to be independent Bernoulli random quantities (or the $y_i(n_i)$ to be independent binomial random quantities) conditional on random quantities θ_i, $i = 1, 2$;

(ii) (θ_1, θ_2) are assigned a joint probability distribution Q;

(iii) by the strong law of large numbers, $\theta_i = \lim_{n_i \to \infty}(y_i(n_i)/n_i)$, so that Q may be interpreted as "joint beliefs about the limiting relative frequencies of 1's in the two sequences".

The model is completed by the specification of $dQ(\theta_1, \theta_2)$, whose detailed form will, of course, depend on the particular beliefs appropriate to the actual practical application of the model. At a qualitative level, we note the following possibilities:

(a) knowledge of the limiting relative frequency for one of the sequences would not change beliefs about outcomes in the other sequence, so that we have the independent form of prior specification, $dQ(\theta_1, \theta_2) = dQ(\theta_1)dQ(\theta_2)$;

(b) the limiting relative frequency for the second sequence will necessarily be greater than that for the first sequence (due, for example, to a known improvement in a drug or an electronic component under test), so that $dQ(\theta_1, \theta_2)$ is zero outside the range $0 \leq \theta_1 < \theta_2 \leq 1$;

(c) there is a real possibility, to which an individual assigns probability π, say, that, in fact, the limiting frequencies could turn out to be equal, so that, writing $\theta = \theta_1 = \theta_2$, in this case $dQ(\theta_1, \theta_2)$ has the form

$$\pi dQ^*(\theta) + (1 - \pi)dQ^+(\theta_1, \theta_2)$$

and the representation, for (y_{1n_1}, y_{2n_2}), say, has the form

$$p(y_1(n_1), y_2(n_2)) = \pi \int_0^1 \mathrm{Bi}(y_1(n_1) \mid n_1, \theta)\, \mathrm{Bi}(y_2(n_2) \mid n_2, \theta)\, dQ^*(\theta)$$
$$+ (1 - \pi) \int_0^1 \int_0^1 \mathrm{Bi}(y_1(n_1) \mid n_1, \theta_1)\, \mathrm{Bi}(y_2(n_2) \mid n_2, \theta_2)\, dQ^+(\theta_1, \theta_2),$$

where $dQ^+(\theta_1, \theta_2)$ assigns probability over the range of values of (θ_1, θ_2) such that $\theta_1 \neq \theta_2$.

As we shall see later, in Chapter 5, the general form of representation of beliefs for observables defined in terms of the two sequences, together with detailed specifications of $dQ(\theta_1, \theta_2)$, enables us to explore coherently any desired aspect of the learning process. For example, we may have observed that out of the first n_1, n_2

patients receiving drug treatments 1, 2, respectively, $y_1(n_1)$ and $y_2(n_2)$ survived, and, on the basis of this information, wish to make judgements about the relative performance of the drugs were they to be used on a large future sequence of patients. This might be done by calculating, for example,

$$p(\lim_{N\to\infty}(y_1(N)/N) - \lim_{N\to\infty}(y_2(N)/N) \mid y_1(n_1), y_2(n_2)),$$

which, in the language of the conventional paradigm, is the "posterior density for $\theta_1 - \theta_2$, given $y_1(n_1)$, $y_2(n_2)$".

Clearly, the discussion and resulting forms of representation which we have given for the case of unrestrictedly exchangeable sequences of 0–1 random quantities can be extended to more general cases. One possible generalisation of Definition 4.13 is the following.

Definition 4.14. (***Unrestricted exchangeability for sequences with predictive sufficient statistics***). *Sequences of random quantities $x_{i1}, x_{i2}, \ldots$ taking values in $X_i, i = 1, \ldots, m$, are said to be unrestrictedly infinitely exchangeable if each sequence is infinitely exchangeable and, in addition, for all $n_i \leq N_i$, $i = 1, \ldots, m$,*

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m) \mid \boldsymbol{t}_{N_1}, \ldots, \boldsymbol{t}_{N_m}) = \prod_{i=1}^{m} p(\boldsymbol{x}_i(n_i) \mid \boldsymbol{t}_{N_i})$$

where $\boldsymbol{t}_{N_i} = \boldsymbol{t}_{N_i}(\boldsymbol{x}_i(N_i))$, $i = 1, \ldots, m$, are separately predictive sufficient statistics for the individual sequences.

In general, given m unrestrictedly exchangeable sequences of random quantities, $x_{i1}, x_{i2}, \ldots$, with x_{ij} taking values in X_i, we typically arrive at a representation of the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta^*} \prod_{i=1}^{m} \prod_{j=1}^{n_i} p_i(x_{ij} \mid \theta_i)\, dQ(\theta_1, \ldots, \theta_m),$$

where $\Theta^* = \prod_{i=1}^{m} \Theta_i$ and the parametric families

$$p_i(x \mid \theta_i), \quad x \in X_i, \quad \theta_i \in \Theta_i, \quad i = 1, \ldots, m,$$

have been identified through consideration of sufficient statistics of fixed dimension, or whatever, as discussed in previous sections. Most often, the fact that the k sequences are being considered together will mean that the random quantities $x_{i1}, x_{i2}, \ldots$ relate to the same form of measurement or counting procedure for all $i = 1, \ldots, m$, so that typically we will have $p_i(x \mid \theta_i) = p(x \mid \theta,), i = 1, \ldots, m$, where the parameters correspond to strong law limits of functions of the sufficient statistics. The following forms are frequently assumed in applications.

Example 4.9. ***(Binomial).*** If $y_i(n_i)$ denotes the number of 1's in the first n_i outcomes of the ith of m unrestrictedly exchangeable sequences of 0 – 1 random quantities, then

$$p(y_1(n_1), \ldots, y_m(n_m)) = \int_{[0,1]^m} \prod_{i=1}^{m} \mathrm{Bi}(y_i(n_i) \mid \theta_i, n_i)\, dQ(\theta_1, \ldots, \theta_m).$$

Example 4.10. ***(Multinomial).*** If $\boldsymbol{y}_i(n_i)$ denotes the category membership count (into the first k of $k+1$ exclusive categories) from the first n_i outcomes of the ith of m unrestrictedly exchangeable sequences of "0 – 1 random vectors" (see Section 4.3), then

$$p(\boldsymbol{y}_1(n_1), \ldots, \boldsymbol{y}_m(n_m)) = \int_{\Theta^m} \prod_{i=1}^{m} \mathrm{Mu}_k(\boldsymbol{y}_i(n_i) \mid \theta_i, n_i)\, dQ(\theta_1, \ldots, \theta_m),$$

where $\theta_i = \lim_{n \to \infty}(\boldsymbol{y}_i(n)/n)$ and $\Theta = \{\theta = (\theta_1, \ldots, \theta_k)$ such that $0 \leq \theta_i \leq 1, 1 \leq i \leq k,$ and $\theta_1 + \cdots + \theta_k \leq 1\}$. This model describes beliefs about an $m \times (k+1)$ *contingency table* of count data, with row totals $n_1, \ldots, n_m$. It generalises the case of the $m \times 2$ contingency table described in Example 4.9. ◁

Example 4.11. ***(Normal).*** If $x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m$, denote real-valued observations from m unrestrictedly exchangeable sequences of real-valued random quantities, the assumed sufficiency of the sample sum and sum of squares within each sequence might lead to the representation

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta^m} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \mathrm{N}(x_{ij} \mid \mu_i, \lambda_i)\, dQ(\theta).$$

where, with $\bar{x}_n(i) = n^{-1}(x_{i1} + \cdots + x_{in})$ and $s_n^2(i) = n^{-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_n(i))^2$, we have $\mu_i = \lim_{n \to \infty} \bar{x}_n(i)$, $\lambda_i^{-1} = \lim_{n \to \infty} s_n^2(i)$, $\theta = (\mu_1, \ldots, \mu_m, \lambda_1, \ldots, \lambda_m)$ and $\Theta = \Re^m \times (\Re^+)^m$.

In many applications, the further judgement is made that $\lambda_1 = \cdots = \lambda_m = \lambda$, say, so that the representation then takes the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Re^m \times \Re^+} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \mathrm{N}(x_{ij} \mid \mu_i, \lambda)\, dQ(\mu_1, \ldots, \mu_m, \lambda).$$

This is the model most often used to describe beliefs about a *one-way layout* of measurement data. ◁

As in the case of 0 – 1 random quantities with $m = 2$, discussed earlier in this section, we could make analogous remarks concerning the various qualitative forms of specification of the prior distribution Q that might be made in these cases. We shall not pursue this further here, but will comment further in Section 4.7.5.

4.6.3 Structured Layouts

Let us now consider the situation described in (ii) of Section 4.6.1, where the random quantity x_{ijk} is triple-subscripted to indicate that it is the kth of K "replicates" of an observable in "context" $i \in I$, subject to "treatment" $j \in J$. In general terms, we have a *two-way layout*, having I rows and J columns, with K replicates in each of the IJ cells.

In such contexts, most individuals would find it unacceptable to make a judgement of complete exchangeability for the random quantities x_{ijk}. For example, if rows represent age-groups, columns correspond to different drug treatments, replicates refer to sequences of patients within each age-group/treatment combination and the x_{ijk} measure death-rates, say, it is typically not the case that beliefs about the x_{ijk} would be invariant under permutations of the subscript i. On the other hand, for the kinds of mechanisms routinely used to allocate patients to treatment groups in clinical trials, many individuals would have exchangeable beliefs about the sequence $x_{ij1}, x_{ij2}, \ldots$ for any fixed i, j.

Technically, such a situation corresponds to the invariance of joint beliefs for the collection of random quantities, x_{ijk}, under some restricted set of permutations of the subscripts, rather than under the unrestricted set of all possible permutations (which would correspond to complete exchangeability). The precise nature of the appropriate set of invariances encapsulating beliefs in a particular application will, of course, depend on the actual perceived partial exchangeabilities in that application. In what follows, we shall simply motivate, using very minimal exchangeability assumptions, a model which is widely used in the context of the two-way layout. There is no suggestion that the particular form discussed has any special status, or *ought* to be routinely adopted, or whatever.

Suppose that, for any fixed i, j, we think of $x_{ij1}, x_{ij2}, \ldots$ as a (potentially) infinite sequence of real-valued random quantities ($x \in \Re$), such that the IJ sequences of this kind, with I and J fixed, are judged to be unrestrictedly exchangeable. If further assumptions of centred spherical symmetry or sufficiency for each sequence then lead to the normal form of representation, we have

$$p(x_{11}(n_{11}), \ldots, x_{IJ}(n_{IJ})) = \int_{\Theta^{IJ}} \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{n_{ij}} \mathrm{N}(x_{ijk} \mid \mu_{ij}, \lambda_{ij})\, dQ(\theta),$$

where $\theta = (\mu_{11}, \ldots, \mu_{IJ}, \lambda_{11}, \ldots, \lambda_{IJ})$ and $\Theta = \Re^{IJ} \times (\Re^{+})^{IJ}$, so that conditional, for each (i, j), on the strong law limits

$$\mu_{ij} = \lim_{K \to \infty} K^{-1}(x_{ij1} + \cdots + x_{ijK}) = \lim_{K \to \infty} K^{-1} \bar{x}_{ij}(K),$$

$$(\lambda_{ij})^{-1} = \lim_{K \to \infty} K^{-1} \sum_{k=1}^{K} (x_{ijk} - \bar{x}_{ij}(K))^2 = \lim_{K \to \infty} s_{ij}^2(K),$$

the x_{ijk} are assumed independently and normally distributed with means μ_{ij} and variances $(\lambda_{ij})^{-1}$.

In many cases, the nature of the observational process leads to the judgement that $\lim_{K\to\infty} s_{ij}^2(K)$ may be assumed to be the same for all (i, j), so that $\lambda_{ij} = \lambda$, say, for all i, j. Letting

$$\mu_{i\bullet} = \lim_{K\to\infty} K^{-1}J^{-1}\sum_{j=1}^{J} \bar{x}_{ij}(K) = J^{-1}\sum_{j=1}^{J} \mu_{ij}$$

$$\mu_{\bullet j} = \lim_{K\to\infty} K^{-1}I^{-1}\sum_{i=1}^{I} \bar{x}_{ij}(K) = I^{-1}\sum_{i=1}^{I} \mu_{ij}$$

$$\mu_{\bullet\bullet} = \lim_{K\to\infty} K^{-1}I^{-1}J^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J} \bar{x}_{ij}(K) = I^{-1}\sum_{i=1}^{I} \mu_{i\bullet} = J^{-1}\sum_{j=1}^{J} \mu_{\bullet j}$$

denote the strong law limits of the row averages, column averages and overall average, respectively, from the two-way layout with I and J fixed, we can always write

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where

$$\alpha_i = (\mu_{i\bullet} - \mu), \quad \beta_j = (\mu_{\bullet j} - \mu), \quad \gamma_{ij} = (\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j}),$$

so that the random quantities x_{ijk} are conditionally independently distributed with

$$p(x_{ijk} \mid \mu, \alpha_i, \beta_j, \gamma_{ij}, \lambda) = \mathrm{N}(x_{ijk} \mid \mu + \alpha_i + \beta_j + \gamma_{ij}, \lambda).$$

The full model representation is then completed by the specification of a prior distribution Q for λ and any IJ linearly independent combinations of the μ_{ij}. In conventional terminology, μ is referred to as the *overall mean*, α_i as the ith *row effect*, β_j as the jth *column effect* and γ_{ij} as the (ij)th *interaction effect*. Collectively, the $\{\alpha_i\}$ and $\{\beta_j\}$ are referred to as the *main effects* and $\{\gamma_{ij}\}$ as the *interactions*. Interest in applications often centres on whether or not interactions or main effects are close to zero and, if not, on making inferences about the magnitudes of differences between different row or column effects.

In the above discussion, our exchangeability assumptions were restricted to the sequence $x_{ij1}, x_{ij2}, \ldots$ for fixed i, j. It is possible, of course, that further forms of symmetric beliefs might be judged reasonable for certain permutations of the i, j subscripts. We shall return to this possibility in Section 4.6.5, where we shall see that certain further assumptions of invariance lead naturally to the idea of hierarchical representations of beliefs.

4.6.4 Covariates

In (iii) of Section 4.6, we gave examples of situations where beliefs about sequences of observables $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$ are functionally dependent, in some sense, on the observed values, $z_i, i = 1, \ldots, m$, of a related sequence of (random) quantities. We shall refer to the latter as *covariates* and, in recognition of this dependency, we shall denote the joint density of $x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m$, by

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m) \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_m).$$

The examples which follow illustrate some of the typical forms assumed in applications. Again, there is no suggestion that these particular forms have any special status; they simply illustrate some of the kinds of models which are commonly used.

Example 4.12. ***(Bioassay).*** Suppose that at each of m specified dose levels, $z_1, \ldots, z_m$, of a toxic substance, typically measured on a logarithmic scale, sequences of 0 – 1 random quantities, $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$, are to be observed, where $x_{ij} = 1$ if the jth animal receiving dose z_i survives, $x_{ij} = 0$ otherwise. If, for each $i = 1, \ldots, m$, the sequences $x_{i1}, x_{i2}, \ldots$ are judged exchangeable, and if we denote the number of survivors out of n_i animals observed in the ith sequence by $y_i(n_i) = x_{i1} + \cdots + x_{in_i}$, a straightforward generalisation of the corollary to Proposition 4.18 implies a representation of the form

$$p(y_1(n_1), \ldots, y_m(n_m) \mid \boldsymbol{z}) = \int_{[0,1]^m} \prod_{i=1}^{m} \mathrm{Bi}\Big(y_i(n_i) \mid \theta_i(\boldsymbol{z}), n_i\Big)\, dQ(\theta(\boldsymbol{z})),$$

where $\boldsymbol{z} = (z_1, \ldots, z_m)$, $\theta(\boldsymbol{z}) = (\theta_1(\boldsymbol{z}), \ldots, \theta_m(\boldsymbol{z}))$ and $\theta_i(\boldsymbol{z}) = \lim_{n \to \infty} n^{-1} y_i(n)$.

In many situations, investigators often find it reasonable to assume that

$$\theta_i(\boldsymbol{z}) = \theta(z_i) = G(\phi; z_i),$$

where the functional form G (usually monotone increasing from 0 to 1) is specified, but ϕ is a random quantity. Functions having the form $G(\phi; z_i) = G(\phi_1 + \phi_2 z_i)$, with $\phi_1 \in \Re, \phi_2 \in \Re^+$, are widely used (see, for example, Hewlett and Plackett, 1979), with

$$G(\phi_1 + \phi_2 z_i) = \int_{-\infty}^{\phi_1 + \phi_2 z_i} \mathrm{N}(\mu \mid 0, 1) d\mu \quad \text{(the } \textit{probit} \text{ model)}$$

and

$$G(\phi_1 + \phi_2 z_i) = \exp(\phi_1 + \phi_2 z_i)/\{1 + \exp(\phi_1 + \phi_2 z_i)\} \quad \text{(the } \textit{logit} \text{ model)}$$

being the most common. For any specified $G(.; z_i)$, the required representation has the form

$$p(y_1(n_1), \ldots, y_m(n_m) \mid \boldsymbol{z}) = \int_{\Phi} \prod_{i=1}^{m} \mathrm{Bi}\Big(y_i(n_i) \mid G(\phi; z_i), n_i\Big) dQ^*(\phi),$$

with $dQ^*(\phi)$ specifying a prior distribution for $\phi \in \Phi$. In practice, the specification of Q might be facilitated by reparametrising from ϕ to a more suitable (1-1) transformation $\Psi = \Psi(\phi)$. In the probit and logit cases, for example, $\psi_1 = -\phi_1/\phi_2$ corresponds to the (log) dose, z_i, at which $G(\phi_1 + \phi_2 z_i) = 1/2$. Beliefs about ψ_1 then correspond to beliefs about the (log-) dose level for which the survival frequency in a large series of animals would equal 1/2, the so-called LD50 dose. Experimenters might typically be more accustomed to thinking in terms of $(-\phi_1/\phi_2, \phi_2)$, say, than in terms of (ϕ_1, ϕ_2). ◁

Example 4.13. ***(Growth-curves).*** Suppose that at each of m specified time points, say $z_1, \ldots, z_m$, sequences of real-valued random quantities, $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$, are to be observed, where x_{ij} is the jth replicate measurement (perhaps on a logarithmic scale) of the size or weight of the subject or object of interest at time z_i. Suppose further that the kinds of judgements outlined in Example 4.11 are made about the sequences $x_{i1}, x_{i2}, \ldots$, with $i = 1, \ldots, m$, so that we have the representation

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta_z} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \mathrm{N}(x_{ij} \mid \mu_i(\boldsymbol{z}), \lambda_i(\boldsymbol{z}))\, dQ(\boldsymbol{\theta}(\boldsymbol{z})),$$

where $\boldsymbol{\theta}(\boldsymbol{z}) = (\mu_1(\boldsymbol{z}), \ldots, \mu_m(\boldsymbol{z}), \lambda_1(\boldsymbol{z}), \ldots, \lambda_m(\boldsymbol{z}))$ and $\Theta_z = \Re^m \times (\Re^+)^m$.

In many such situations, the judgement is made that $\lambda_1(\boldsymbol{z}) = \cdots = \lambda_m(\boldsymbol{z}) = \lambda$ (particularly if measurements are made on a logarithmic scale) and that

$$\mu_i(\boldsymbol{z}) = \mu_i(z_i) = g(\boldsymbol{\phi}; z_i),$$

where the functional form g (usually monotone increasing) is specified, but $\boldsymbol{\phi}$ is a random quantity. Commonly assumed forms include

$$g(\boldsymbol{\phi}; z_i) = (\phi_1 + \phi_2 \phi_3^{z_i})^{-1}, \quad \text{(the } \textit{logistic} \text{ model)}$$

and

$$g(\boldsymbol{\phi}; z_i) = \phi_1 + \phi_2 z_i \quad \text{(the } \textit{straight-line} \text{ model)}.$$

For any specified $g(\cdot; z_i)$, the joint predictive density representation has the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta_z} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \mathrm{N}\big(x_{ij} \mid g(\boldsymbol{\phi}, z_i), \lambda\big)\, dQ(\boldsymbol{\phi}, \lambda),$$

where $dQ(\boldsymbol{\phi}, \lambda)$ specifying a prior distribution for $\boldsymbol{\phi} \in \Phi$ and $\lambda \in \Re^+$.

As with Example 4.12, specification of Q might be facilitated if we reparametrise from ϕ to a more suitable (1-1) transformation, $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\phi})$. In the logistic case, for example, we might take $\psi_1 = \phi_1^{-1}$, corresponding to the "saturation" growth level reached as $z_i \to \infty$, and $\psi_2 = (\phi_1 + \phi_2)^{-1}$, corresponding to the growth level at the "time origin", $z_i = 0$. Beliefs about ψ_1, ψ_2 then acquire an operational meaning as beliefs about the average growth-levels, at times "∞" and "0", respectively, that would be observed from a large number of replicate measurements. A third possible parameter to which investigators could easily relate in some applications might be $\psi_3 = \log[\phi_1\phi_2/(2\phi_1 + \phi_2)]/\log(\phi_3)$, the time at which growth is half-way from the initial to the final level. ◁

Example 4.14. ***(Multiple regression).*** Suppose that, for each $i = 1, \ldots, m$, sequences of real-valued random quantities $x_{i1}, x_{i2}, \ldots$ are to be observed, where each x_{ij} is related to certain specified observed quantities $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})$ and judgements are made which lead to the belief representation

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta} \prod_{i=1}^{m} \prod_{j=1}^{n_i} N(x_{ij} \mid \mu_i(\boldsymbol{z}_i), \lambda_i(\boldsymbol{z}_i))\, dQ(\boldsymbol{\theta}(\boldsymbol{z})).$$

where

$$\mu_i(\boldsymbol{z}_i) = \lim_{n \to \infty} \bar{x}_n(i), \quad \lambda_i^{-1}(\boldsymbol{z}_i) = \lim_{n \to \infty} s_n^2(i).$$

$$\boldsymbol{\theta}(\boldsymbol{z}) = (\mu_1(\boldsymbol{z}_1), \ldots, \mu_m(\boldsymbol{z}_m), \quad \lambda_1(\boldsymbol{z}_1), \ldots, \lambda_m(\boldsymbol{z}_m)$$

and $\Theta = \Re^m \times (\Re^+)^m$, with $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$.

In many situations, the further judgements are made that $\lambda_i(\boldsymbol{z}_i) = \lambda_i = \lambda$ and $\mu_i(\boldsymbol{z}_i) = \mu(\boldsymbol{z}_i), i = 1, \ldots, m$, where λ and $\mu(.)$ are unknown, but the latter is assumed to be a "smooth" function, adequately approximated by a first-order Taylor expansion, so that, for some (unspecified) $\boldsymbol{z}^*$,

$$\mu(\boldsymbol{z}_i) = \mu(\boldsymbol{z}^*) + (\boldsymbol{z}_i - \boldsymbol{z}^*) \bigtriangledown \mu(\boldsymbol{z}^*) = \boldsymbol{a}_i \boldsymbol{\theta},$$

where we define

$$\boldsymbol{a}_i = (1, z_{i1}, \ldots, z_{ik}) \text{ (row vector)}$$

and

$$\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_k)' \text{ (column vector)}$$

with

$$\theta_0 = \mu(\boldsymbol{z}^*) - \boldsymbol{z}^* \bigtriangledown \mu(\boldsymbol{z}^*), \theta_i = [\bigtriangledown \mu(\boldsymbol{z}^*)]_i, i = 1, \ldots, k.$$

Conditional on $\boldsymbol{\phi} = (\boldsymbol{\theta}, \lambda)$, the joint distribution of

$$\boldsymbol{x} = (\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m))$$

is thus seen to be multivariate normal, $N_n(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{\lambda})$, where $\boldsymbol{A}$ is an $n \times k$ matrix ($n = n_1 + \cdots + n_m$), whose rows consist of $\boldsymbol{a}_1$ replicated n_1 times, followed by $\boldsymbol{a}_2$ replicated n_2 times, and so on, and $\boldsymbol{\lambda} = \lambda \boldsymbol{I}_n$, with $\boldsymbol{I}_n$ denoting the $n \times n$ identity matrix. The unconditional representation can therefore be written as

$$p(\boldsymbol{x}) = \int_{\Re^{k+1} \times \Re^+} N_n(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{\lambda})\, dQ(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

It is conventional to refer to $z_{1j}, z_{2j}, \ldots$ as values of the *regressor variables* $z^{(j)}, j = 1, \ldots, k$, to $\boldsymbol{\theta}$ as the vector of *regression coefficients* and $\boldsymbol{A}$ as the *design matrix*. The form $\mu(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{\theta}$ is called a *regression equation* and the structure

$$E(\boldsymbol{x} \mid \boldsymbol{A}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{A}\boldsymbol{\theta}$$

is said to define a *linear model*. If $k = 1$, we have the *simple regression* (*straight-line*) model, $E(x_{ij}) = \theta_0 + \theta_1 z_{i1}$; for $k \geq 2$, we have a *multiple regression* model.

From an operational point of view, beliefs about $\boldsymbol{\theta}$ in the general case relate to beliefs about the intercept (θ_0) of the regression equation and the marginal rates of change ($\theta_1, \ldots, \theta_k$) of the x_{ij} with respect to the regressor variables ($z_1, \ldots, z_k$). However, within this general structure we can represent various special cases such as $z^{(j)} = z^j$ (*polynomial regression*) or $z^{(j)} = \sin(jH/N)$, for some N (a version of *trigonometric regression*); in these cases, beliefs about $\boldsymbol{\theta}$ will stem from rather different considerations. ◁

Specification of the kinds of structures which we have illustrated in Examples 4.12 to 4.14 essentially reduces to the same process as we have seen in earlier representations of joint predictive densities as integral mixtures. We proceed *as if*:

(i) the random quantities are *conditionally independent*, given the values of the relevant *covariates*, $\boldsymbol{z}$, and given the *unknown parameters*, $\boldsymbol{\phi}$;
(ii) the latter are assigned a prior distribution, $dQ(\boldsymbol{\phi})$.

In many cases, the likelihood, defined through conditional independence, involves familiar probability models, often of exponential family form (as with the binomial, normal and multivariate examples seen above), but with at least some of the usual "labelling" parameters replaced by more complex functional forms involving the covariates. From a conceptual point of view, this is all that really needs to be said for the time being. However, when we consider the applications of such models, together with the problems of computation, approximation, etc., which arise in implementing the Bayesian learning process, it is often useful to have a more structured taxonomy in mind: for example, linear versus non-linear functional forms; normal versus non-normal distributions, and so on.

4.6.5 Hierarchical Models

In Section 4.6.2, we considered the general situation where several sequences of random quantities, $x_{i1}, x_{i2}, \ldots, i = 1, \ldots, m$ are judged unrestrictedly infinitely exchangeable, leading typically to a joint density representation of the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\Theta^m} \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(x_{ij} \mid \theta_i)\, dQ(\theta_1, \ldots, \theta_m).$$

We remarked at that time that nothing can be said, *in general*, about the prior specification $Q(\theta_1, \ldots, \theta_m)$, since this must reflect whatever beliefs are appropriate for the specific application being modelled. However, it is often the case that additional judgements about relationships among the m sequences lead to interestingly structured forms of $Q(\theta_1, \ldots, \theta_m)$.

In Section 4.6.1, we noted some of the possible contexts in which judgements of exchangeability might be appropriate not only for the random quantities *within* each of m separate sequence of observables, but also *between* the m strong law limits of appropriately defined statistics for each of the sequences. The following examples illustrate this kind of structured judgement and the forms of *hierarchical model* which result.

Example 4.15. ***(Exchangeable binomial parameters).*** Suppose that we have unrestrictedly infinitely exchangeable sequences of 0–1 random quantities, $x_{i1}, x_{i2}, \ldots$, with $i = 1, \ldots, m$. Then, for $i = 1, 2, \ldots$, $[n_i, y_i(n_i) = x_{i1} + \cdots + x_{in_i}]$, is a sufficient statistic for the ith sequence and

$$p(y_1(n_1), \ldots, y_m(n_m)) = \int_{[0,1]^m} p(y_1(n_1), \ldots, y_m(n_m) \mid \theta_1, \ldots, \theta_m) dQ(Q_1, \ldots, \theta_n)$$
$$= \int_{[0,1]^m} \prod_{i=1}^{m} \mathrm{Bi}(y_i(n_i) \mid \theta_i, n_i)\, dQ(\theta_1, \ldots, \theta_m),$$

where

$$\theta_i = \lim_{n \to \infty} (y_i(n)/n).$$

As we remarked in Section 4.6.1, if the sequences consists of success-failure outcomes on repeated trials with m different (but, to all intents and purposes, "similar") types of component, it might be reasonable to judge the m "long-run success frequencies" to be themselves exchangeable. This corresponds to specifying *an exchangeable form of prior distribution for the parameters* $\theta_1, \ldots, \theta_m$. If the m types of component can be thought of as a selection from a potentially infinite sequence of similar components, we then have (see Section 4.3.3) the general representation

$$Q(\theta_1, \ldots, \theta_m) = \int_{\mathfrak{I}} Q(\theta_1, \ldots, \theta_m \mid G)\, d\Pi(G)$$
$$= \int_{\mathfrak{I}} \prod_{i=1}^{m} G(\theta_i)\, d\Pi(G).$$

The complete model structure is then seen to have the *hierarchical form*

$$p(y_1(n_1), \ldots, y_m(n_m) \mid \theta_1, \ldots, \theta_m) = \prod_{i=1}^{m} \mathrm{Bi}(y_i(n_i) \mid \theta_i, n_i)$$

$$Q(\theta_1, \ldots, \theta_m \mid G) = \prod_{i=1}^{m} G(\theta_i)$$

$$\Pi(G)$$

In conventional terminology, the first stage of the hierarchy relates data to parameters via binomial distributions; the second stage models the binomial parameters *as if* they were a random sample from a distribution G; the third, and final, stage specifies beliefs about G. ◁

The above example is readily generalised to the case of exchangeable parameters for any one-parameter exponential family. In practice, beliefs about G might concentrate on a particular parametric family, so that, assuming the existence of the appropriate densities, the prior specification takes the form

$$g(\theta_1, \ldots, \theta_m) = \int_\Phi g(\theta_1, \ldots, \theta_m \mid \phi)\, d\Pi(\phi) = \int_\Phi \prod_{i=1}^{m} g(\theta_i \mid \phi)\, d\Pi(\phi)$$

and, for appropriate sufficient statistics $y_i(n_i)$, $i = 1, \ldots, m$, defines the hierarchical structure

$$p(y_1(n_1), \ldots, y_m(n_m) \mid \theta_1, \ldots, \theta_m) = \prod_{i=1}^{m} p(y_i(n_i) \mid \theta_i)$$

$$g(\theta_1, \ldots, \theta_m \mid \phi) = \prod_{i=1}^{m} g(\theta_i \mid \phi)$$

$$\Pi(\phi).$$

As before, the first stage of the hierarchy relates data to parameters in a form assumed to be independent of G; the second stage now models the parameters *as if* they were a random sample from a parametric family labelled by the *hyperparameter* $\phi \in \Phi$; the third, and final, stage specifies beliefs about the hyperparameter. Such beliefs acquire operational significance by identifying the hyperparameter with appropriate strong law limits of observables, as we shall indicate in the following example.

Example 4.16. ***(Exchangeable normal mean parameters).*** Suppose that we have m unrestrictedly infinitely exchangeable sequences $x_{i1}, x_{i2}, \ldots$, $i = 1, \ldots, m$, of real valued random quantities, for which (see Example 4.11) the joint density has the representation

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m)) = \int_{\mathbb{R}^m \cdot \mathbb{R}^+} \prod_{i=1}^{m} \prod_{j=1}^{n_i} \mathrm{N}(x_{ij} \mid \mu_i, \lambda)\, dQ(\mu_1, \ldots, \mu_m, \lambda),$$

where we recall that $\lambda^{-1} = \lim_{n \to \infty} s_n^2(i)$ and $\mu_i = \lim_{n \to \infty} \bar{x}_n(i)$, where

$$n\bar{x}_n(i) = (x_{i1} + \cdots + x_{in}), \quad ns_n^2(i) = \sum_{j=1}^{n} (x_{ij} - \bar{x}_n(i))^2, \; i = 1, \ldots, m.$$

So far as the specification of $Q(\mu_1, \ldots, \mu_m, \lambda)$ is concerned, we first note that in many applications it is helpful to think in terms of

$$Q(\mu_1, \ldots, \mu_m, \lambda) = Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda) Q_\lambda(\lambda).$$

for some Q_μ, Q_λ. In some cases, knowledge of the strong law limits of sums of squares about the mean may be judged irrelevant to the assessment of beliefs for strong law limits of the sample averages: in such cases, $Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda)$ will not depend on λ. In other cases, we might believe, for example, that variation among the limiting sample averages is certainly bigger (or certainly smaller) than within-sequence variation of observations about the sample mean: in such cases, $Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda)$ will involve λ. In either case, it is useful to think in terms of the product form of Q.

Now suppose that, conditional on λ, the limiting sample means are judged exchangeable. If the m sequences can be thought of as a selection from a potentially infinite collection of similar sequences, we have (see Section 4.3) a further representation of Q_μ in the form

$$Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda) = \int_{\mathfrak{I}} Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda, G)\, d\Pi(G \mid \lambda)$$
$$= \int_{\mathfrak{I}} \prod_{i=1}^{m} G(\mu_i \mid \lambda)\, d\Pi(G \mid \lambda).$$

The complete model then has the hierarchical structure

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m) \mid \mu_1, \ldots, \mu_m, \lambda) = \prod_{i=1}^{m} p(\boldsymbol{x}_i(n_i) \mid \mu_i, \lambda)$$

$$Q_\mu(\mu_1, \ldots, \mu_m \mid \lambda, G) = \prod_{i=1}^{m} G(\mu_i \mid \lambda)$$

$$\Pi(G \mid \lambda) Q_\lambda(\lambda).$$

In practice, beliefs about G, given λ, might concentrate on a particular parametric family, so that, assuming the existence of the appropriate densities, the hierarchical structure would take the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m) \mid \mu_1, \ldots, \mu_m, \lambda) = \prod_{i=1}^{m} p(\boldsymbol{x}_i(n_i) \mid \mu_i, \lambda)$$

$$g_\mu(\mu_1, \ldots, \mu_m \mid \lambda, \phi) = \prod_{i=1}^{m} g_\mu(\mu_i \mid \lambda, \phi)$$

$$\Pi x(\phi \mid \lambda)\, Q_\lambda(\lambda).$$

For an explicit example of this, suppose that, given a potentially infinite sequence $\mu_1, \mu_2, \ldots$ (or, more concretely, $\bar{x}_{n1}(1), \bar{x}_{n2}(2), \ldots$, for very large $n_1, n_2, \ldots$) the quantities $m, \bar{\mu}(m) = m^{-1}(\mu_1 + \cdots + \mu_m)$ and $s^2(m) = m^{-1} \sum_{j=1}^{m} (\mu_j - \bar{\mu}(m))^2$ (or the large sample analogues of $\bar{\mu}(m)$ and $s^2(m)$) were judged sufficient for the sequence. It would then be natural (see Section 4.5) to take $g_\mu(\mu_i \mid \lambda, \phi) = N(\mu_i \mid \phi_1, \phi_2)$, where

$$\phi_1 = \lim_{m \to \infty} \bar{\mu}(m), \quad \phi_2 = \lim_{m \to \infty} s^2(m).$$

From an operational standpoint, the final stage specification of the joint prior distribution for ϕ_1, ϕ_2 and λ then reduces to a specification of beliefs about the following limits of observable quantities (for large m and $n_i, i = 1, \ldots, m$):

(i) the mean of all the observations from all the sequences (ϕ_1);

(ii) the mean sum of squares of the individual sequence means about the overall mean (ϕ_2);

(iii) the mean (over sequences) of the mean sum of squares of observations within a sequence about the sequence mean (λ).

The precise form of specification at this stage will, of course, depend on the particular situation in which the model is being applied. ◁

Hierarchical modelling provides a powerful and flexible approach to the representation of beliefs about observables in extended data structures, and is being increasingly used in statistical modelling and analysis. This section has merely provided a brief introduction to the basic ideas and the way such structures arise naturally within a subjectivist, modelling framework. In the context of the Bayesian learning process, further brief discussion will be given in Section 5.6.4, where links will be made with *empirical Bayes* ideas.

An extensive discussion of hierarchical modelling will be given in the volumes *Bayesian Computation* and *Bayesian Methods*. A selection of references to the literature on inference for hierarchical models will be given in Section 5.6.4.

4.7 PRAGMATIC ASPECTS

4.7.1 Finite and Infinite Exchangeability

The de Finetti representation theorem for 0-1 random quantities, and the various extensions we have been considering in this chapter, characterise forms of $p(x_1, \ldots, x_n)$ for observables $x_1, \ldots, x_n$ assumed to be part of an *infinite* exchangeable sequence. However, in general, mathematical representations which correspond to probabilistic mixing over conditionally independent parametric forms do not hold for *finite* exchangeable sequences.

To see this, consider $n = 2$ and finitely exchangeable 0-1 x_1, x_2, such that

$$p(x_1 = 0, x_2 = 0) = p(x_1 = 1, x_2 = 1) = 0$$
$$p(x_1 = 1, x_2 = 0) = p(x_1 = 0, x_2 = 1) = \tfrac{1}{2}.$$

If the de Finetti representation held, we would have

$$\int_0^1 \theta^2 dQ(\theta) = \int_0^1 (1-\theta)^2 dQ(\theta) = 0,$$

for some $Q(\theta)$, an impossibility since the latter would have to assign probability one to both $\theta = 0$ and $\theta = 1$ (Diaconis and Freedman, 1980a).

It appears, therefore, that there is a potential conflict between realistic modelling (acknowledging the necessarily finite nature of actual exchangeability judgements) and the use of conventional mathematical representations (derived on the basis of assumed infinite exchangeability).

To discuss this problem, let us call an exchangeable sequence, $x_1, \ldots, x_n$, with $x_i \in X$, *N-extendible* if it is part of the longer exchangeable sequence $x_1, \ldots, x_N$. Practical judgements of exchangeability for specific observables $x_1, \ldots, x_n$ are typically of this kind: the $x_1, \ldots, x_n$ can be considered as part of a larger, *but finite*, potential sequence of exchangeable observables. Infinite exchangeability corresponds to the possibly unrealistic assumption of N-extendibility for all $N > n$.

In general, the assumption of infinite exchangeability implies that the probability assigned to an event $(x_1, \ldots, x_n) \in E \subseteq X^n$ is of the form

$$P_Q(E) = \int F^n(E) dQ(F),$$

for some Q. If we denote by $P(E)$ the corresponding probability assigned under N-extendibility for a specific N, a possible measure of the "distortion" introduced by assuming infinite exchangeability is given by

$$\sup_E |P(E) - P_Q(E)|,$$

where the supremum is taken over all events in the appropriate σ-field on X^n. Intuitively, one might feel that if $x_1, \ldots, x_n$ is N-extendible for some $N >> n$, the "distortion" should be somewhat negligible. This is made precise by the following.

Proposition 4.19. (***Finite approximation of infinite exchangeability***). *With the preceding notation, there exists Q such that*

$$\sup_E |P(E) - P_Q(E)| \leq \frac{f(n)n}{N},$$

where $f(n)$ is the number of elements in X, if the latter is finite, and $f(n) = (n-1)$ otherwise.

Proof. See Diaconis and Freedman (1980a) for a rigorous statement and technical details. ◁

The message is clear and somewhat comforting. If a realistic judgement of N-extendibility for large, but finite, N is replaced by the mathematically convenient assumption of infinite exchangeability, no important distortion will occur in quantifying uncertainties.

For further discussion, see Diaconis (1977), Jaynes (1986) and Hill (1992). For extensions of Proposition 4.19 to multivariate and linear model structures, see Diaconis *et al.* (1992).

4.7.2 Parametric and Nonparametric Models

In Section 4.3, we saw that the assumption of exchangeability for a sequence $x_1, x_2, \ldots$ of real-valued random quantities implied a general representation of the joint distribution function of $x_1, \ldots, x_n$ of the form

$$P(x_1, \ldots, x_n) = \int_{\mathfrak{F}} \prod_{i=1}^{n} F(x_i) dQ(F),$$

where

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

and F_n is the empirical distribution function defined by $x_1, \ldots, x_n$. This implies that we should proceed *as if* we have a random sample from an unknown distribution function F, with Q representing our beliefs about "what the empirical distribution would look like for large n".

As we remarked at the end of Section 4.3.3, the task of assessing and representing such a belief distribution Q over the set $\mathfrak{F}$ of all possible distribution functions is by no means straightforward, since F is, effectively, an infinite-dimensional parameter. Most of this chapter has therefore been devoted to exploring additional features of beliefs which justify the restriction of $\mathfrak{F}$ to families of distributions having explicit mathematical forms involving only a finite-dimensional labelling parameter.

Conventionally, albeit somewhat paradoxically, representations in the finite-dimensional case are referred to as *parametric models*, whereas those involving the infinite-dimensional parameter are referred to as *nonparametric models*! The technical key to Bayesian nonparametric modelling is thus seen to be the specification of appropriate probability measures over function spaces, rather than over finite-dimensional real spaces, as in the parametric case. For this reason, the Bayesian analysis of nonparametric models requires considerably more mathematical machinery than the corresponding analysis of parametric models. In the rest of this volume we will deal exclusively with the parametric case, postponing a treatment of nonparametric problems to the volumes *Bayesian Computation* and *Bayesian Methods*.

Among important references on this topic, we note Whittle (1958), Hill (1968, 1988, 1992), Dickey (1969), Kimeldorf and Wahba (1970), Good and Gaskins (1971, 1980), Ferguson (1973, 1974), Leonard (1973), Antoniak (1974), Doksum (1974), Susarla and van Ryzin (1976), Ferguson and Phadia (1979), Dalal and Hall (1980), Dykstra and Laud (1981), Padgett and Wei (1981), Rolin (1983), Lo (1984), Thorburn (1986), Kestemont (1987), Berliner and Hill (1988), Wahba (1988), Hjort (1990), Lenk (1991) and Lavine (1992a).

As we have seen, the use of specific parametric forms can often be given a *formal* motivation or justification as the coherent representation of certain forms of

belief characterised by invariance or sufficiency properties. In practice, of course, there are often less formal, more *pragmatic*, reasons for choosing to work with a particular parametric model (as there often are for acting, formally, *as if* particular forms of summary statistic were sufficient!). In particular, specific parametric models are often suggested by *exploratory data analysis* (typically involving graphical techniques to identify plausible distributional shapes and forms of relationship with covariates), or by *experience* (i.e., historical reference to "similar" situations, where a given model seemed "to work") or by *scientific theory* (which determines that a specific mathematical relationship "must" hold, in accordance with an assumed "law"). In each case, of course, the choice involves subjective judgements; for example, regarding such things as the "straightness" of a graphical normal plot, the "similarity" between a current and a previous trial, and the "applicability of a theory to the situation under study. From the standpoint of the general representation theorem, such judgements correspond to acting *as if* one has a Q which concentrates on a subset of $\Im$ defined in terms of a finite-dimensional labelling parameter.

4.7.3 Model Elaboration

However, 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). It would always be prudent, therefore, to "expand one's consciousness" a little in relation to an intended model in order to review the judgements that have been made. Depending on the context, the following kinds of critical questions might be appropriate:

(i) is it reasonable to assume that all the observables form a "homogeneous sample", or might a few of them be "aberrant" in some sense?

(ii) is it reasonable to apply the modelling assumptions to the observables on their original scale of measurement, or should the scale be transformed to logarithms, reciprocals, or whatever?

(iii) when considering temporally or spatially related observables, is it reasonable to have made a particular conditional independence assumption, or should some form of correlation be taken into account?

(iv) if some, but not all, potential covariates have been included in the model, is it reasonable to have excluded the others, or might some of them be important, either individually or in conjunction with covariates already included?

We shall consider each of these possibilities in turn, indicating briefly the kinds of elaboration of the "first thought of" model that might be considered.

Outlier elaboration. Suppose that judgements about a sequence $x_1, x_2, \ldots$ of real-valued random quantities have led to serious consideration of the model

$$p(x_1, \ldots, x_n) = \int_{\Re \times \Re^+} \prod_{i=1}^{n} \mathrm{N}(x_i \mid \mu, \lambda)\, dQ(\mu, \lambda),$$

but, on reflection, it is thought wise to allow for the fact that (an unknown) one of $x_1, \ldots, x_n$ *might* be aberrant. If aberrant observations are assumed to be such that a sequence of them would have a limiting mean equal to μ, but a limiting mean square about the mean equal to $(\gamma\lambda)^{-1}$, $0 < \gamma < 1$, where μ and λ^{-1} denote the corresponding limits for non-aberrant observations, a suitable form of elaborated model might be

$$p(x_1, \ldots, x_n) = \pi \int_{\Re \times \Re^+} \prod_{i=1}^{n} \mathrm{N}(x_i \mid \mu, \lambda)\, dQ(\mu, \lambda)$$

$$+ (1 - \pi) \int_{\Re \times \Re^+ \times [0,1]} \sum_{j=1}^{n} \frac{1}{n} \mathrm{N}(x_j \mid \mu, \gamma\lambda) \left\{ \prod_{i \neq j} \mathrm{N}(x_i \mid \mu, \lambda) \right\} dQ(\mu, \lambda)\, dQ(\gamma).$$

This model corresponds to an initial assumption that, with specified probability π, there are no aberrant observations, but, with probability $1 - \pi$, there is precisely one aberrant observation, which is equally likely to be any one of $x_1, \ldots, x_n$. Generalisations to cover more than one possible aberrant observation can be constructed in an obviously analogous manner. Such models are usually referred to as "outlier" models, since $\gamma < 1$ implies an increased probability that, in the observed sample $x_1, \ldots, x_n$, the aberrant observation will "outlie". Since for an aberrant observation x, $\mathrm{E}[(x - \mu)^2 \mid \mu, \lambda, \gamma] = (\gamma\lambda)^{-1}$, prior belief in the relative inaccuracy of an aberrant observation as a "predictor" of μ is reflected in the weight attached by the prior distribution $Q(\gamma)$ to values of γ much smaller than 1.

De Finetti (1961) and Box and Tiao (1968) are pioneering Bayesian papers on this topic. More recent literature includes: Dawid (1973), O'Hagan (1979, 1988b, 1990), Freeman (1980), Smith (1983), West (1984, 1985), Pettit and Smith (1985), Arnaiz and Ruíz-Rivas (1986), Muirhead (1986), Pettit (1986, 1992), Guttman and Peña (1988) and Peña and Guttman (1993).

Transformation elaboration. Suppose now that judgements about a sequence $x_1, x_2, \ldots$ of real-valued random quantities are such that it seems reasonable to suppose that, *if a suitable* γ *were identified*, beliefs about the sequence $x_1^{(\gamma)}, x_2^{(\gamma)}, \ldots$ defined by

$$x_i^{(\gamma)} = (x_i^{\gamma} - 1)/\gamma \qquad (\gamma \neq 0, \gamma \in \Gamma)$$
$$= \log(x_i) \qquad (\gamma = 0),$$

would plausibly have the representation

$$p(x_1^{(\gamma)}, \ldots, x_n^{(\gamma)}) = \int_{\Re \times \Re^+} \prod_{i=1}^{n} \mathrm{N}(x_i^{(\gamma)} \mid \mu, \lambda) dQ^*(\mu, \lambda \mid \gamma).$$

It then follows that

$$p(x_1, \ldots, x_n) = \int_{\Re \times \Re^+ \times \Gamma} \prod_{i=1}^{n} \mathrm{N}(x_1^{(\gamma)} \mid \mu, \lambda) J(\boldsymbol{x}, \gamma)\, dQ^*(\mu, \lambda \mid \gamma)\, dQ^+(\gamma)$$

where

$$J(\boldsymbol{x}, \gamma) = \prod_{i=1}^{n} [dx_i^{(\gamma)}/dx_i].$$

The case $\gamma = 1$ corresponds to assuming a normal parametric model for the observations on their original scale of measurement. If Γ includes values such as $\gamma = -1, \gamma = 1/2, \gamma = 0$, the elaborated model admits the possibility that transformations such as reciprocal, square root, or logarithm, might provide a better scale on which to assume a normal parametric model. Judgements about the relative plausibilities of these and other possible transformations are then incorporated in Q^+. For detailed developments see Box and Cox (1964), Pericchi (1981) and Sweeting (1984, 1985).

Correlation elaboration. Suppose that judgements about $x_1, x_2, \ldots$ again lead to a "first thought of model in which

$$p(x_1, \ldots, x_n \mid \mu, \lambda) = \prod_{i=1}^{n} \mathrm{N}(x_i \mid \mu, \lambda),$$

but that it is then recognised that there may be a serial correlation structure among $x_1, \ldots, x_n$ (since, for example, the observations correspond to successive time-points, $t = 1, t = 2$, etc.) A possible extension of the representation to incorporate such correlation might be to assume that, for a given $\gamma \in [-1, 1)$, and conditional on μ and λ, the correlation between x_t and x_{t+h} is given by $R(x_t, x_{t+h} \mid \mu, \lambda, \gamma) = \gamma^h$, so that

$$p(\boldsymbol{x} \mid \mu, \lambda, \gamma) = p(x_1, \ldots, x_n \mid \mu, \lambda, \gamma) = \mathrm{N}_n(\boldsymbol{x} \mid \mu \mathbf{1}, \lambda \Gamma^{-1}),$$

where

$$\Gamma = \begin{bmatrix} 1 & \gamma & \gamma^2 & \ldots & \gamma^{n-1} \\ \gamma & 1 & \gamma & & \gamma^{n-2} \\ \vdots & & \ldots & & \vdots \\ \gamma^{n-1} & \gamma^{n-2} & \gamma^{n-3} & \ldots & 1 \end{bmatrix}$$

The elaborated model then becomes, for some Q^*, Q^+,

$$p(x_1, \ldots, x_n) = \int_{\Re \times \Re^+ \times (-1,1)} \mathrm{N}_n(\boldsymbol{x} \mid \mu \mathbf{1}, \lambda \Gamma^{-1})\, dQ^*(\mu, \lambda \mid \gamma)\, dQ^+(\gamma)$$

The "first thought of" model corresponds to $\gamma = 0$ and beliefs about the relative plausibility of this value compared with other possible values of positive or negative correlation are reflected in the specification of Q^+.

Covariate elaboration. Suppose that the "first thought of" model for the observables $\boldsymbol{x} = (\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_m(n_m))$, where $\boldsymbol{x}_i(n_i) = (x_{i1}, \ldots, x_{in_i})$ denotes replicate observations corresponding to the observed value $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})$ of the covariates $z_1, \ldots, z_k$, is the multiple regression model with representation

$$p(\boldsymbol{x}) = \int_{\Re^{k+1} \times \Re^+} \mathrm{N}_n(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{\theta}, \lambda)\, dQ(\boldsymbol{\theta}, \lambda)$$

as described in Example 4.16 of Section 4.6. If it is subsequently thought that covariates $z^{k+1}, \ldots, z^l$ should also have been taken into account, a suitable elaboration might take the form of an extended regression model

$$p(\boldsymbol{x}) = \int_{\Re^{l+1} \times \Re^+} \mathrm{N}_n(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\gamma}, \lambda)\, dQ^*(\boldsymbol{\theta}, \boldsymbol{\gamma}, \lambda),$$

where $\boldsymbol{B}$ consists of rows containing $\boldsymbol{b}_i = (z_{ik+1}, \ldots, z_{il})$ replicated n_i times, $i = 1, \ldots, m$ and $\boldsymbol{\gamma} = (\theta_{k+1}, \ldots, \theta_l)$ denotes the regression coefficients of the additional regressor variables $z_{k+1}, \ldots, z_l$. The value $\boldsymbol{\gamma} = \mathbf{0}$ corresponds to the "first thought of" model.

In all these cases, an initially considered representation of the form

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\phi})\, dQ(\boldsymbol{\phi})$$

is replaced by an elaborated representation

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{\phi}, \boldsymbol{\gamma})\, dQ^*(\boldsymbol{\phi}, \boldsymbol{\gamma}),$$

the latter reducing to the original representation on setting the elaboration parameter $\boldsymbol{\gamma}$ equal to $\mathbf{0}$. Inference about such a $\boldsymbol{\gamma}$, imaginatively chosen to reflect interesting possible forms of departure from the original model, often provides a natural basis for checking on the adequacy of an initially proposed model, as well as learning about the directions in which the model needs extending.

Other Bayesian approaches to the problem of covariate selection include Bernardo and Bermúdez (1985), Mitchell and Beauchamp (1988) and George and McCulloch (1993a).

4.7.4 Model Simplification

The process of model elaboration, outlined in the previous section, consists in expanding a "first thought of" model to include additional parameters (and possibly covariates), reflecting features of the situation whose omission from the original model formulation is, on reflection, thought to be possibly injudicious.

The process of model simplification is, in a sense, the converse. In reviewing a currently proposed model, we might wonder whether some parameters (or covariates) have been unnecessary included, in the sense that a simpler form of model might be perfectly adequate. As it stands, of course, this latter consideration is somewhat ill-defined: the "adequacy", or otherwise, of a particular form of belief representation can only be judged in relation to the consequence arising from actions taken on the basis of such beliefs. These and other questions relating to the fundamentally important area of model comparison and model choice will be considered at length in Chapter 6. For the present, it will suffice just to give an indication of some particular forms of model simplification that are routinely considered.

Equality of parameters. In Section 4.6, we analysed the situation where several sequences of observables are judged unrestrictedly infinitely exchangeable, leading to a general representation of the form

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_n(n_m)) = \int_{\Theta^*} \prod_{i=1}^{m} \prod_{j=1}^{n_i} p(x_{ij} \mid \theta_i)\, dQ(\theta_1, \ldots, \theta_m),$$

where $\theta_i \in \Theta_i, \Theta^* = \prod_{i=1}^{m} \Theta_i$ and the parameter θ_i relating to the ith sequence can typically be interpreted as the limit of a suitable summary statistic for the ith sequence. If, on the other hand, the simplifying judgement were made that, in fact, the labelling of the sequences is irrelevant and that any combined collection of observables from any or all of the sequences would be completely exchangeable, we would have the representation

$$p(\boldsymbol{x}_1(n_1), \ldots, \boldsymbol{x}_n(n_m)) = \int_{\Theta^*} \prod_{i=1}^{m} \prod_{j=1}^{n} p(x_{ij} \mid \theta)\, dQ(\theta)$$

where the same parameter $\theta \in \Theta$ now suffices to label the parametric model for each of the sequences. In conventional terminology, the simplified representation is sometimes referred to as the *null-hypothesis* ($\theta_1 = \cdots = \theta_m$) and the original representation as the *alternative hypothesis* ($\theta_1 \neq \cdots \neq \theta_m$). As we saw in Section 4.6 (for the case of two 0-1 sequences), rather than opt for sure for one or other of these representations, we could take a *mixture* of the two (with weight π, say, on the null representation and $1 - \pi$ on the alternative, general, representation). This

form of representation will be considered in more detail in Chapter 6, where it will be shown to provide a possible basis for evaluating the relative plausibility of the "null and alternative hypotheses" in the light of data.

Absence of effects. In Section 4.6, we considered the situation of a structured layout with replicate sequences of observations in each of IJ cells, and a possible parametric model representation involving *row effects* $(\alpha_1, \ldots, \alpha_I)$, *column effects* $(\beta_1, \ldots, \beta_J)$ and *interaction effects* $(\gamma_{11}, \ldots, \gamma_{IJ})$. A commonly considered simplifying assumption is that there are no interaction effects $(\gamma_{11} = \cdots = \gamma_{IJ} = 0)$, so that large sample means in individual cells are just the additive combination of the corresponding large sample row and column means.

Further possible simplifying judgements would be that the row (or column) labelling is irrelevant, so that $\alpha_1 = \cdots = \alpha_I = 0$ (or $\beta_1 = \cdots = \beta_J = 0$) and large sample cell means coincide with column (or row) means. Again, conventional terminology would refer to these simplifying judgements as "null hypotheses".

Omission of covariates. Considering, for example, the multiple regression case, described in Example 4.14 of Section 4.6 and reconsidered in the previous section on model elaboration, we see that here the simplification process is very clearly just the converse of the elaboration process. If γ denotes the regression coefficients of the covariates we are considering omitting, then the model corresponding to $\gamma = 0$ provides the required simplification.

In fact, in all the cases of elaboration which we considered in the previous section, setting the "elaboration parameter" γ to 0 provides a natural form of simplification of potential interest. Whether the process of model comparison and choice is seen as one of elaboration or of simplification is then very much a pragmatic issue of whether we begin with a "smaller" model and consider making it "bigger", or vice versa. In any case, issues of model comparison and choice require a separate detailed and extensive treatment, which we defer until Chapter 6.

4.7.5 Prior Distributions

The operational subjectivist approach to modelling views predictive models as representations of beliefs about observables (including limiting, large-sample functions of observables, conventionally referred to as parameters). Invariance and sufficiency considerations have then been shown to justify a structured approach to predictive models in terms of integral mixtures of parametric models with respect to distributions for the labelling parameters. In familiar terminology, we specify a distribution for the observables conditional on unknown parameters (a *sampling distribution*, defining a *likelihood*), together with a distribution for the unknown parameters (a *prior distribution*). *It is the combination of prior and likelihood which defines the overall model.* In terms of the mixture representation, the specification of a prior distribution for unknown parameters is therefore an essential

and unavoidable part of the process of representing beliefs about observables and hence of learning from experience.

> From the operational, subjectivist perspective, it is meaningless to approach modelling solely in terms of the parametric component and ignoring the prior distribution. We are, therefore, in fundamental disagreement with approaches to statistical modelling and analysis which proceed only on the basis of the sampling distribution or likelihood and treat the prior distribution as something optional, irrelevant, or even subversive (see Appendix B).

That said, it should be readily acknowledged that the process of representing prior beliefs itself involves a number of both conceptual and practical difficulties, and certainly cannot be summarily dealt with in a superficial or glib manner.

From a conceptual point of view, as we have repeatedly stressed throughout this chapter, prior beliefs about parameters typically acquire an operational significance and interpretation as beliefs about limiting (large-sample) functions of observables. Care must therefore obviously be taken to ensure that prior specifications respect logical or other constraints pertaining to such limits. Often, the specification process will be facilitated by suitable "reparametrisation".

From a practical point of view, detailed treatment of specific cases is very much a matter of "methods" rather than "theory" and will be dealt with in the third volume of this series. However, a general overview of representation strategies, together with a number of illustrative examples, will be given in the inference context in Chapter 5. In particular, we shall see that the range of creative possibilities opened up by the consideration of mixtures, asymptotics, robustness and sensitivity analysis, as well as novel and flexible forms of inference reporting, provides a rich and illuminating perspective and framework for inference, within which many of the apparent difficulties associated with the precise specification of prior distributions are seen to be of far less significance than is commonly asserted by critics of the Bayesian approach.

4.8 DISCUSSION AND FURTHER REFERENCES

4.8.1 Representation Theorems

The original representation theorem for exchangeable 0 – 1 random quantities appears in de Finetti (1930), the concept of exchangeability having been considered earlier by Haag (1924) and also in the early 1930's by Khintchine (1932). Extensions to the case of general exchangeable random quantities appear in de Finetti (1937/1964) and Dynkin (1953), with an abstract analytical version appearing in Hewitt and Savage (1955). Seminal extensions to more complex forms of symmetry (partial exchangeability) can be found in de Finetti (1938) and Freedman

(1962). See Diaconis and Freedman (1980b) and Wechsler (1993) for overviews and generalisations of the concept of exchangeability.

Recent and current developments have generated an extensive catalogue of characterisations of distributions via both invariance and sufficiency conditions. Important progress is made in Diaconis and Freedman (1984, 1987, 1990) and Küchler and Lauritzen (1989). See, also, Ressel (1985). Useful reviews are given by Aldous (1985), Diaconis (1988a) and, from a rather different perspective, Lauritzen (1982, 1988). The conference proceedings edited by Koch and Spizzichino (1982) also provides a wealth of related material and references. For related developments from a reliability perspective, see Barlow and Mendel (1992, 1994) and Mendel (1992).

4.8.2 Subjectivity and Objectivity

Our approach to modelling has been dictated by a subjectivist, operational concern with individual beliefs about (potential) observables. Through judgements of symmetry, partial symmetry, more complex invariance or sufficiency, we have seen how mixtures over conditionally independent "parameter-labelled" forms arise as typical representations of such beliefs. We have noted how this illuminates, and puts into perspective, linguistic separation into "likelihood" (or "sampling model") and "prior" components. But we have also stressed that, from our standpoint, the two are actually inseparable in defining a belief model.

In contrast, traditional discussion of a statistical model typically refers to the parametric form as "the model". The latter then defines "objective" probabilities for outcomes defined in terms of observables, these probabilities being determined by the values of the "unknown parameters". It is often implicit in such discussion that if the "true" parameter were known, the corresponding parametric form would be the "true" model for the observables. Clearly, such an approach seeks to make a very clear distinction between the nature of observables and parameters. It is as if, given the "true" parameter, the corresponding parametric distribution is seen as part of "objective reality", providing the mechanism whereby the observables are generated. The "prior", on the other hand, is seen as a "subjective" optional extra, a potential contaminant of the objective statements provided by the parametric model.

Clearly, this view has little in common with the approach we have systematically followed in this volume. However, there is an interesting sense, even from our standpoint, in which the parametric model and the prior can be seen as having different roles.

Instead of viewing these roles as corresponding to an objective/subjective dichotomy, we view them in terms of an intersubjective/subjective dichotomy (following Dawid, 1982b, 1986b). To this end, consider a *group* of Bayesians, all concerned with their belief distributions for the same sequence of observables. In the absence of any general agreement over assumptions of symmetry, invariance or

sufficiency, the individuals are each simply left with their own subjective assessments. However, given some set of common assumptions, the results of this chapter imply that the entire group will structure their beliefs using some common form of mixture representation. Within the mixture, the parametric forms adopted will be the same (the *intersubjective* component), while the priors for the parameter will differ from individual to individual (the *subjective* component). Such intersubjective agreement clearly facilitates communication within the group and reduces areas of potential disagreement to just that of different prior judgements for the parameter. As we shall see in Chapter 5, judgements about the parameter will tend more towards a consensus as more data are acquired, so that such a group of Bayesians may eventually come to share very similar beliefs, even if their initial judgements about the parameter were markedly different. We emphasise again, however, that the key element here is intersubjective agreement or consensus. We can find no real role for the idea of objectivity except, perhaps, as a possibly convenient, but potentially dangerously misleading, "shorthand" for intersubjective communality of beliefs.

4.8.3 Critical Issues

We conclude this chapter on modelling with some further comments concerning (i) *The Role and Nature of Models*, (ii) *Structural and Stochastic Assumptions*, (iii) *Identifiability* and (iv) *Robustness Considerations*.

The Role and Nature of Models

In the approach we have adopted, the fundamental notion of a model is that of a predictive probability specification for observables. However, the forms of representation theorems we have been discussing provide, in typical cases, a basis for separating out, if required, two components; the parametric model, and the belief model for the parameters. Indeed, we have drawn attention in Section 4.8.2 to the fact that shared structural belief assumptions among a group of individuals can imply the adoption of a common form of parametric model, while allowing the belief models for the parameters to vary from individual to individual. One might go further and argue that without some element of agreement of this kind there would be great difficulty in obtaining any meaningful form of scientific discussion or possible consensus.

Non-subjectivist discussions of the role and nature of models in statistical analysis tend to have a rather different emphasis (see, for example, Cox, 1990, and Lehmann, 1990). However, such discussions often end up with a similar message, implicit or explicit, about the importance of models in providing a focused framework to serve as a basis for subsequent identification of areas of agreement and disagreement. In order to think about complex phenomena, one must necessarily work with simplified representations. In any given context, there are typically

a number of different choices of degrees of simplification and idealisation that might be adopted and these different choices correspond to what Lehmann calls "a reservoir of models", where

> ... particular emphasis is placed on transparent characterisations or descriptions of the models that would facilitate the understanding of when a given model is appropriate. (Lehmann, 1990)

But appropriate for what? Many authors—including Cox and Lehmann—highlight a distinction between what one might call *scientific* and *technological* approaches to models. The essence of the dichotomy is that scientists are assumed to seek *explanatory* models, which aim at providing insight into and understanding of the "true" mechanisms of the phenomenon under study; whereas technologists are content with *empirical* models, which are not concerned with the "truth", but simply with providing a reliable basis for practical action in predicting and controlling phenomena of interest.

Put very crudely, in terms of our generic notation, explanatory modellers take the form of $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ very seriously, whereas empirical modellers are simply concerned that $p(\boldsymbol{x})$ "works". For an elaboration of the latter view, see Leonard (1980).

The approach we have adopted is compatible with either emphasis. As we have stressed many times, it is observables which provide the touchstone of experience. When comparing rival belief specifications, all other things being equal we are intuitively more impressed with the one which consistently assigns higher probabilities to the things that actually happen. If, in fact, a phenomenon is governed by the specific mechanism $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, a scientist who discovers this and sets $p(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$ will certainly have a $p(\boldsymbol{x})$ that "works".

However, we are personally rather sceptical about taking the science versus technology distinction too seriously. Whilst we would not dispute that there are typically real differences in motivation and rhetoric between scientists and technologists, it seems to us that theories are always ultimately judged by the predictive power they provide. Is there really a meaningful concept of "truth" in this context other than a pragmatic one predicated on $p(\boldsymbol{x})$? We shall return to this issue in Chapter 6, but our prejudices are well-captured in the adage: "*all models are false, but some are useful*".

Structural and Stochastic Assumptions

In Section 4.6, we considered several illustrative examples where, separate from considerations about the complete form of probability specification to be adopted, the key role of the parametric model component $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ was to specify structured forms of expectations for the observables conditional on the parameters. We recall two examples.

In the case of observables x_{ijk} in a two-way layout with replications (Section 4.6.3), with parameters corresponding to overall mean, main effects and interactions, we encountered the form

$$E(x_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij};$$

in the case of a vector of observables $\boldsymbol{x}$ in a multiple regression context with design matrix $\boldsymbol{A}$ (Section 4.6.4, Example 4.14), we encountered the form

$$E(\boldsymbol{x}) = \boldsymbol{A\theta}.$$

In both of these cases, fundamental explanatory or predictive structure is captured by the specification of the conditional expectation, and this aspect can in many cases be thought through separately from the choice of a particular specification of full probability distribution.

Identifiability

A parametric model for which an element of the parametrisation is redundant is said to be non-identified. Such models are often introduced at an early stage of model building (particularly in econometrics) in order to include all parameters which may originally be thought to be relevant. Identifiability is a property of the parametric model, but a Bayesian analysis of a non-identified model is always possible if a proper prior on all the parameters is specified. For detailed discussion of this issue, see Morales (1971), Drèze (1974), Kadane (1974), Florens and Mouchart (1986), Hills (1987) and Florens *et al.* (1990, Section 4.5).

Robustness Considerations

For concreteness, in our earlier discussion of these examples we assumed that the $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ terms were specified in terms of normal distributions. As we demonstrated earlier in this chapter, under the a priori assumption of appropriate invariances, or on the basis of experience with particular applications, such a specification may well be natural and acceptable. However, in many situations the choice of a specific probability distribution may feel a much less "secure" component of the overall modelling process than the choice of conditional expectation structure.

For example, past experience might suggest that departures of observables from assumed expectations resemble a symmetric bell-shaped distribution centred around zero. But a number of families of distributions match these general characteristics, including the normal, Student and logistic families. Faced with a seemingly arbitrary choice, what can be done in a situation like this to obtain further insight and guidance? Does the choice matter? Or are subsequent inferences or predictions robust against such choices?

An exactly analogous problem arises with the choice of mathematical specifications for the prior model component.

In robustness considerations, theoretical analysis—sometimes referred to as "what if?" analysis—has an interesting role to play. Using the inference machinery which we shall develop in Chapter 5, the desired insight and guidance can often be obtained by studying mathematically the ways in which the various "arbitrary" choices affect subsequent forms of inferences and predictions. For example, a "what if?" analysis might consider the effect of a single, aberrant, outlying observation on inferences for main effects in a multiway layout under the alternative assumptions of a normal or Student parametric model distribution. It can be shown that the influence of the aberrant observation is large under the normal assumption, but negligible under the Student assumption, thus providing a potential basis for preferring one or other of the otherwise seemingly arbitrary choices.

More detailed analysis of such robustness issues will be given in Section 5.6.3.