1

Basic ideas and examples

1.1 The statistical problem

Latent variable models provide an important tool for the analysis of multivariate data. They offer a conceptual framework within which many disparate methods can be unified and a base from which new methods can be developed. A statistical model specifies the joint distribution of a set of random variables and it becomes a latent variable model when some of these variables – the latent variables – are unobservable. In a formal sense, therefore, there is nothing special about a latent variable model. The usual apparatus of model-based inference applies, in principle, to all models regardless of their type. The interesting questions concern why latent variables should be introduced into a model in the first place and how their presence contributes to scientific investigation.

One reason, common to many techniques of multivariate analysis, is to reduce dimensionality. If, in some sense, the information contained in the interrelationships of many variables can be conveyed, to a good approximation, in a much smaller set, our ability to 'see' the structure in the data will be much improved. This is the idea which lies behind much of factor analysis and the newer applications of linear structural models. Large-scale statistical enquiries, such as social surveys, generate much more information than can be easily absorbed without drastic summarisation. For example, the questionnaire used in a sample survey may have 50 or 100 questions and replies may be received from 1000 respondents. Elementary statistical methods help to summarise the data by looking at the frequency distributions of responses to individual questions or pairs of questions and by providing summary measures such as percentages and correlation coefficients. However, with so many variables it may still be difficult to see any pattern in their interrelationships. The fact that our ability to visualise relationships is limited to two or three dimensions places us under strong pressure to reduce the dimensionality of the data in a manner which preserves as much of the structure as possible. The reasonableness of such a course is often

Latent Variable Models and Factor Analysis: A Unified Approach, Third Edition.
David Bartholomew, Martin Knott and Irini Moustaki.

evident from the fact that many questions overlap in the sense that they seem to be getting at the same thing. For example, one's views about the desirability of private health care and of tax levels for high earners might both be regarded as a reflection of a basic political position. Indeed, many enquiries are designed to probe such basic attitudes from a variety of angles. The question is then one of how to condense the many variables with which we start into a much smaller number of indices with as little loss of information as possible. Latent variable models provide one way of doing this.

A second reason is that latent quantities figure prominently in many fields to which statistical methods are applied. This is especially true of the social sciences. A cursory inspection of the literature of social research or of public discussion in newspapers or on television will show that much of it centres on entities which are handled as if they were measurable quantities but for which no measuring instrument exists. Business confidence, for example, is spoken of as though it were a real variable, changes in which affect share prices or the value of the currency. Yet business confidence is an ill-defined concept which may be regarded as a convenient shorthand for a whole complex of beliefs and attitudes. The same is true of quality of life, conservatism, and general intelligence. It is virtually impossible to theorise about social phenomena without invoking such hypothetical variables. If such reasoning is to be expressed in the language of mathematics and thus made rigorous, some way must be found of representing such 'quantities' by numbers. The statistician's problem is to establish a theoretical framework within which this can be done. In practice one chooses a variety of indicators which can be measured, such as answers to a set of yes/no questions, and then attempts to extract what is common to them.

In both approaches we arrive at the point where a number of variables have to be summarised. The theoretical approach differs from the pragmatic in that in the former a pre-existing theory directs the search and provides some means of judging the plausibility of any measures which result. We have already spoken of these measures as indices or hypothetical variables. The usual terminology is *latent variables* or *factors*. The term *factor* is so vague as to be almost meaningless, but it is so firmly entrenched in this context that it would be fruitless to try to dislodge it now. We prefer to speak of latent variables since this accurately conveys the idea of something underlying what is observed. However, there is an important distinction to be drawn. In some applications, especially in economics, a latent variable may be real in the sense that it could, in principle at least, be measured. For example, personal wealth is a reasonably well-defined concept which could be expressed in monetary terms, but in practice we may not be able or willing to measure it. Nevertheless we may wish to include it as an explanatory variable in economic models and therefore there is a need to construct some proxy for it from more accessible variables. There will be room for argument about how best to do this, but wide agreement on the existence of the latent variable. In most social applications the latent variables do not have this status. Business confidence is not something which exists in the sense that personal wealth does. It is a summarising concept which comes prior to the indicators of it which we measure. Much of the philosophical debate which takes place on latent variable models centres on *reification*; that is, on speaking as though such things as quality

of life and business confidence were real entities in the sense that length and weight are. However, the usefulness and validity of the methods to be described in this book do not depend primarily on whether one adopts a realist or an instrumentalist view of latent variables. Whether one regards the latent variables as existing in some real world or merely as a means of thinking economically about complex relationships, it is possible to use the methods for prediction or establishing relationships *as if* the theory were dealing with real entities. In fact, as we shall see, some methods, which appear to be purely empirical, lead their users to behave as if they had adopted a latent variable model. We shall return to the question of interpreting latent variables at the end of Chapter 9. In the meantime we note that an interesting discussion of the meaning of a latent variable can be found in Sobel (1994).

1.2 The basic idea

We begin with a very simple example which will be familiar to anyone who has met the notion of spurious correlation in elementary statistics. It concerns the interpretation of a 2×2 contingency table. Suppose that we are presented with Table 1.1. Leaving aside questions of statistical significance, the table exhibits an association between the two variables. If A was being a heavy smoker and B was having lung cancer someone might object that the association was spurious and that it was attributable to some third factor C with which A and B were both associated – such as living in an urban environment. If we go on to look at the association between A and B in the presence and absence of C we might obtain data as set out in Table 1.2. The original association has now vanished and we therefore conclude that the underlying variable C was wholly responsible for it. Although the correlation between the manifest variables might be described as spurious, it is here seen as pointing to an underlying latent variable whose influence we wish to determine.

Even in the absence of any suggestion about C it would still be pertinent to ask whether the original table could be decomposed into two tables exhibiting independence. If so, we might then look at the members of each subgroup to see if they had anything in common, such as most of one group living in an urban environment. The idea can be extended to a p-way table and again we can enquire whether it can be decomposed into sub-tables in which the variables are independent. If this were possible there would be grounds for supposing that there was some latent categorisation which fully explained the original association. The discovery of such a

Table 1.1 A familiar example.

	A	$\bar{A}$	Total
B	350	200	550
$\bar{B}$	150	300	450
	500	500	1000

Table 1.2 Effect of a hidden factor.

	C			$\bar{C}$		
	A	$\bar{A}$	Total	A	$\bar{A}$	Total
B	320	80	400	30	120	150
$\bar{B}$	80	20	100	70	280	350
	400	100	500	100	400	500

decomposition would amount to having found a latent categorical variable for which conditional independence held. The validity of the search does not require the assumption that the goal will be reached. In a similar fashion we can see how two categorical variables might be rendered independent by conditioning on a third continuous latent variable. We now illustrate these rather abstract ideas by showing how they arise with two of the best-known latent variable models.

1.3 Two examples

1.3.1 Binary manifest variables and a single binary latent variable

We now take the argument one step further by introducing a probability model for binary data. In order to do this we shall need to anticipate some of the notation required for the more general treatment given below. Thus suppose there are p binary variables, rather than two as in the last example. Let these be denoted by $x_1, x_2, \ldots, x_p$ with $x_i = 0$ or 1 for all i. Let us consider whether the mutual association of these variables could be accounted for by a single binary variable y. In other words, is it possible to divide the population into two parts so that the xs are mutually independent in each group? It is convenient to label the two hypothetical groups 1 and 0 (as with the xs, any other labelling would serve equally well). The prior distribution of y will be denoted $h(y)$, and this may be written

$$h(1) = P\{y = 1\} = \eta \quad \text{and} \quad h(0) = 1 - h(1). \tag{1.1}$$

The conditional distribution of x_i given y will be that of a Bernoulli random variable written

$$P\{x_i \mid y\} = \pi_{iy}^{x_i}(1 - \pi_{iy})^{1-x_i} \quad (x_i, y = 0, 1), \tag{1.2}$$

where π_{iy} is the probability that $x_i = 1$ when the latent class is y. Notice that in this simple case the form of the distributions h and $P\{x_i \mid y\}$ is not in question; it is only their parameters, η, $\{\pi_{i0}\}$ and $\{\pi_{i1}\}$ which are unspecified by the model.

For this model

$$f(\mathbf{x}) = \eta \prod_{i=1}^{p} \pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i} + (1-\eta)\prod_{i=1}^{p} \pi_{i0}^{x_i}(1-\pi_{i0})^{1-x_i}. \tag{1.3}$$

To test whether such a decomposition is adequate we would fit the probability distribution of (1.3) to the observed frequency distribution of $\mathbf{x}$-vectors and apply a goodness-of-fit test. As we shall see later, the parameters of (1.3) can be estimated by maximum likelihood. If the fit were not adequate we might go on to consider three or more latent classes or, perhaps, to allow y to vary continuously.

If the fit were satisfactory we might wish to have a rule for allocating individuals to one or other of the latent classes on the basis of their $\mathbf{x}$-vector. For this we need the posterior distribution

$$\begin{aligned} h(1\,|\,\mathbf{x}) &= P\{y=1\,|\,\mathbf{x}\} \\ &= \eta\left(\prod_{i=1}^{p}\pi_{i1}^{x_i}(1-\pi_{i1})^{1-x_i}\right)\Big/ f(\mathbf{x}) \\ &= 1\Big/\left[1+\left(\frac{1-\eta}{\eta}\right)\exp\sum_{i=1}^{p}\left\{x_i\,\ln\frac{\pi_{i0}}{\pi_{i1}}+(1-x_i)\,\ln\frac{1-\pi_{i0}}{1-\pi_{i1}}\right\}\right]. \end{aligned} \tag{1.4}$$

Clearly individuals cannot be allocated with certainty, but if estimates of the parameters are available an allocation can be made on the basis of which group is more probable. Thus we could allocate to group 1 if

$$h(1\,|\,\mathbf{x}) > h(0\,|\,\mathbf{x}),$$

that is, if

$$\begin{aligned} X &= \sum_{i=1}^{p} x_i\{\operatorname{logit}\pi_{i1} - \operatorname{logit}\pi_{i0}\} \\ &> \sum_{i=1}^{p}\ln\{(1-\pi_{i0})/(1-\pi_{i1})\} - \operatorname{logit}\eta. \end{aligned} \tag{1.5}$$

where $\operatorname{logit} u = \ln\{u/(1-u)\}$. An interesting feature of this result is that the rule for discrimination depends on the xs in a linear fashion. Here, this is a direct consequence of the fact that the posterior distribution of (1.4) depends on $\mathbf{x}$ only through the linear combination which we may denote by X. In that sense X contains all the relevant information in the data about the latent variable. This is not peculiar to this example but will turn out to be a key idea which is at the heart of the theoretical treatment of Chapter 2.

It is worth emphasising again that much of the arbitrariness in the general approach with which we started has been avoided by fixing the number of latent classes and hence the form of the distribution h. There might, of course, be some prior grounds for expecting two latent groups, but nothing is lost by the assumption because, if it fails, we can go on to try more.

1.3.2 A model based on normal distributions

When $\mathbf{x}$ consists of metrical variables the writing down of a model is a little less straightforward. As before, we might postulate two latent classes and then we should have

$$f(\mathbf{x}) = \eta \prod_{i=1}^{p} g_i(x_i \mid y = 1) + (1 - \eta) \prod_{i=1}^{p} g_i(x_i \mid y = 0), \tag{1.6}$$

where $g_i(x_i \mid .)$ denotes the conditional density of x_i given the particular value of y. However, we are now faced with the choice of conditional distributions for x_i. There is now no natural choice as there was when the xs were binary. We could, of course, make a plausible guess, fit the resulting model and try to justify our choice retrospectively by a goodness-of-fit test. Thus if a normal conditional distribution seemed reasonable we could proceed along the same lines as in Section 1.3.1. Models constructed in this way will be discussed in Chapter 6.

1.4 A broader theoretical view

Having introduced the basic idea of a latent variable model, we are now ready to move on to the general case where the latent variables may not be determined in either number or form. As our primary concern is with the logic of the general model we shall treat all variables as continuous, but this is purely for notational simplicity and does not affect the key idea.

There are two sorts of variables to be considered and they will be distinguished as follows. Variables which can be directly observed, also known as *manifest* variables, will be denoted by x. A collection of p manifest variables will be distinguished by subscripts and written as a column vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$. In the interests of notational economy we shall not usually distinguish between random variables and the values which they take. When necessary, the distinction will be effected by the addition of a second subscript, thus x_{ih} will be the observed value of random variable x_i for the hth sample member and $\mathbf{x}_h$ will be that member's $\mathbf{x}$-vector. The corresponding notation for latent variables will be y and q, and such variables will form the column vector $\mathbf{y}$. In practice we shall be concerned with the case where q is much smaller than p. Since both manifest and latent variables, by definition, vary from one individual to another they are represented in the theory by random variables. The relationships between them must therefore be expressed in terms of probability distributions, so, for example, after the xs have been observed the information we

have about $\mathbf{y}$ is contained in its conditional distribution given $\mathbf{x}$. Although we are expressing the theory in terms of continuous variables, the modifications required for the other combinations of Table 1.3 on page 11 are straightforward and do not bear upon the main points to be made.

As only $\mathbf{x}$ can be observed, any inference must be based on their joint distribution whose density may be expressed as

$$f(\mathbf{x}) = \int_{R_y} h(\mathbf{y}) g(\mathbf{x} \mid \mathbf{y}) d\mathbf{y}, \tag{1.7}$$

where $h(\mathbf{y})$ is the prior distribution of $\mathbf{y}$, $g(\mathbf{x} \mid \mathbf{y})$ is the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ and R_y is the range space of $\mathbf{y}$ (this will usually be omitted). Our main interest is in what can be known about $\mathbf{y}$ after $\mathbf{x}$ has been observed. This information is wholly conveyed by the conditional density $h(\mathbf{y} \mid \mathbf{x})$, deduced from Bayes' theorem,

$$h(\mathbf{y} \mid \mathbf{x}) = h(\mathbf{y}) g(\mathbf{x} \mid \mathbf{y}) / f(\mathbf{x}). \tag{1.8}$$

We use h for both the prior distribution and the conditional distribution, but which one is meant is always clear from the notation. The nature of the problem we face is now clear. In order to find $h(\mathbf{y} \mid \mathbf{x})$ we need to know both h and g, but all that we can estimate is f. It is obvious that h and g are not uniquely determined by (1.7) and thus, at this level of generality, we cannot obtain a complete specification of $h(\mathbf{y} \mid \mathbf{x})$. For further progress to be made we must place some further restriction on the classes of functions to be considered. In fact (1.7) and (1.8) do not specify a model, they are merely different ways of expressing the fact that $\mathbf{x}$ and $\mathbf{y}$ are random variables that are mutually dependent on one another. No other assumption is involved. However, rather more is implied in our discussion than we have yet brought out. If the xs are each related to one or more of the ys then there will be correlations among the xs. Thus if x_1 and x_2 both depend on y_1 we may expect the common influence of y_1 to induce a correlation between x_1 and x_2. Conversely if x_1 and x_2 were uncorrelated there would be no grounds for supposing that they had anything in common. Taking this one step further, if x_1 and x_2 are uncorrelated *when* y_1 *is held fixed* we may infer that no other y is needed to account for their relationship since the existence of such a y would induce a correlation even if y_1 were fixed.

In general we are saying that if the dependencies among the xs are induced by a set of latent variables $\mathbf{y}$ then when all ys are accounted for, the xs will be independent if all the ys are held fixed. If this were not so the set of ys would not be *complete* and we should have to add at least one more. Thus q must be chosen so that

$$g(\mathbf{x} \mid \mathbf{y}) = \prod_{i=1}^{p} g_i(x_i \mid \mathbf{y}). \tag{1.9}$$

This is often spoken of as the assumption (or axiom) of conditional (or local) independence. But it is misleading to think of it as an assumption of the kind that could be tested empirically because there is no way in which $\mathbf{y}$ can be fixed and therefore

no way in which the independence can be tested. It is better regarded as a definition of what we mean when we say that the set of latent variables $\mathbf{y}$ is complete. In other words, that $\mathbf{y}$ is sufficient to explain the dependencies among the xs. We are asking whether $f(\mathbf{x})$ admits the representation

$$f(\mathbf{x}) = \int h(\mathbf{y}) \prod_{i=1}^{p} g_i(x_i \mid \mathbf{y}) d\mathbf{y}, \tag{1.10}$$

for some q, h and $\{g_i\}$. In practice we are interested in whether (1.10) is an adequate representation for some small value of q. The dependence of (1.10) on q is concealed by the notation and is thus easily overlooked. We do not *assume* that (1.10) holds; a key part of our analysis is directed to discovering the smallest q for which such a representation is adequate.

The treatment we have just given is both general and abstract. In the following chapter we shall propose a family of conditional distributions to take the place of $g_i(x_i \mid \mathbf{y})$ which will meet most practical needs. However, there are several points which the foregoing treatment makes very clear which have often been overlooked when we come down to particulars. For example, once $f(\mathbf{x})$ is known, or estimated, we are not free to choose h and g independently. Our choice is constrained by the need for (1.7) to be satisfied. Thus if we want to talk of 'estimating' the prior distribution $h(\mathbf{y})$, as is sometimes done, such an estimate will be constrained by the choice already made for g. Similarly, any attempt to locate individuals in the latent space using the conditional distribution $h(\mathbf{y} \mid \mathbf{x})$ must recognise the essential arbitrariness of the prior distribution $h(\mathbf{y})$. As we shall see, this indeterminacy is central to understanding what is often called the *factor scores* problem.

A more subtle point concerns the interpretation of the prior distribution, $h(y)$ itself (we temporarily restrict the discussion to the one-dimensional case). The latent variable is essentially a construct and therefore there is no need for it to exist in the ordinary sense of that word. Consequently, its distribution does not exist either and it is therefore meaningless to speak of estimating it!

1.5 Illustration of an alternative approach

In the foregoing development we constructed the model from its basic ingredients, finally arriving at the joint distribution which we can actually observe. We might try to go in the other direction starting from what we observe, namely $f(\mathbf{x})$, and deducing what the ingredients would have to be if $f(\mathbf{x})$ is to be the end-product. Suppose, for example, our sample of $\mathbf{x}$s could be regarded as coming from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. We might then ask whether the multivariate normal distribution admits a representation of the form (1.10) and, if so, whether it is unique. It is easy to find one such representation using standard results of distribution theory. Suppose, for example, that

$$\mathbf{y} \sim N_q(\mathbf{0}, \mathbf{I})$$

and

$$\mathbf{x} \mid \mathbf{y} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y}, \boldsymbol{\Psi}), \tag{1.11}$$

where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of coefficients and $\boldsymbol{\Psi}$ is a diagonal matrix of variances. It then follows that

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}), \tag{1.12}$$

which is of the required form. Note that although this representation works for all $q \leq p$ there is no implication in general that a $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ can be found such that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is equal to the given $\boldsymbol{\Sigma}$. Every model specified by (1.11) leads to a multivariate normal $\mathbf{x}$, but if $q < p$ the converse is not true. The point of the argument is to show that the model (1.11) is worth entertaining if the xs have a multivariate normal distribution.

The posterior distribution of $\mathbf{y}$ is easily obtained by standard methods and it, too, turns out to be normal. Thus

$$\mathbf{y} \mid \mathbf{x} \sim N_q\left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}\right), \tag{1.13}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ and $q < p$. The mean of this distribution might then be used to predict $\mathbf{y}$ for a given $\mathbf{x}$ and the precision of the predictions would be given by the elements of the covariance matrix.

Unfortunately the decomposition of (1.11) is not unique as we now show. In the traditional approach to factor analysis this feature is met in connection with *rotation*, but the point is a general one which applies whether or not $\mathbf{x}$ is normal.

Suppose that $\mathbf{y}$ is continuous and that we make a one-to-one transformation of the factor space from $\mathbf{y}$ to $\mathbf{v}$. This will have no effect on $f(\mathbf{x})$ since it is merely a change of variable in the integral (1.7), but both of the functions h and g will be changed. In the case of h the *form* of the prior distribution will, in general, be different, and in the case of g there will be a change, for example, in the regression of $\mathbf{x}$ on the latent variables. It is thus clear that there is no unique way of expressing $f(\mathbf{x})$ as in (1.10) and therefore no empirical means of distinguishing among the possibilities. We are thus not entitled to draw conclusions from any analysis which would be vitiated by a transformation of the latent space. However, there may be some representations which are easier to interpret than others. We note in the present case, from (1.11), that the regression of x_i on $\mathbf{y}$ is linear and this enables us to interpret the elements of $\boldsymbol{\Lambda}$ as weights determining the effect of each $\mathbf{y}$ on a particular x_i. Any non-linear transformation of $\mathbf{y}$ would destroy this relationship.

Another way of looking at the matter is to argue that the indeterminacy of h leaves us free to adopt a metric for $\mathbf{y}$ such that h has some convenient form. A normal scale is familiar so we might require each y_j to have a standard normal distribution. If, as a further convenience, we make the ys independent as in (1.11) then the form of g_i is uniquely determined and we would then note that we had the additional benefit of

linear regressions. This choice is essentially a matter of *calibration*; we are deciding on the properties we wish our scale to have.

In general, if we find the representation (1.10) is possible, we may fix either h or $\{g_i\}$; in the normal case either approach leads us to (1.11).

If $\mathbf{x}$ is normal there is an important transformation which leaves the form of h unchanged and which thus still leaves a degree of arbitrariness about $\{g_i\}$. This is the rotation which we referred to above. Suppose $q \geq 2$; then the orthogonal transformation $\mathbf{v} = \mathbf{M}\mathbf{y}$ ($\mathbf{M}'\mathbf{M} = \mathbf{I}$) gives

$$\mathbf{v} \sim N_q(\mathbf{0}, \mathbf{I}),$$

which is the same distribution as $\mathbf{y}$ had. The conditional distribution is now

$$\mathbf{x} \mid \mathbf{v} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{M}'\mathbf{v}, \boldsymbol{\Psi}), \tag{1.14}$$

so that a model with weights $\boldsymbol{\Lambda}$ is indistinguishable from one with weights $\boldsymbol{\Lambda}\mathbf{M}'$. The effect of orthogonally transforming the latent space is thus exactly the same as transforming the weight matrix. The joint distribution of $\mathbf{x}$ is, of course, unaffected by this. In the one case the covariance matrix is $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ and in the other it is $\boldsymbol{\Lambda}\mathbf{M}'\mathbf{M}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, and these are equal because $\mathbf{M}'\mathbf{M} = \mathbf{I}$.

The indeterminacy of the factor model revealed by this analysis has advantages and disadvantages. So far as determining q, the dimensionality of the latent space, is concerned, there is no problem. But from a purely statistical point of view the arbitrariness is unsatisfactory and in Chapter 3 we shall consider how it might be removed. However, there may be practical advantages in allowing the analyst freedom to choose from among a set of transformations that which has most substantive meaning. This too is a matter to which we shall return.

The reader already familiar with factor analysis will have recognised many of the formulae in this section, even though the setting may be unfamiliar. The usual treatment, to which we shall come later, starts with the linear regression implicit in (1.11) and then adds the distributional assumptions in a convenient but more or less arbitrary way. In particular, the essential role of the conditional independence postulate is thereby obscured. The advantage of starting with the distribution of $\mathbf{x}$ is that it leads to the usual model in a more compelling way but, at the same time, makes the essential arbitrariness of some of the usual assumptions clearer. We shall return to these points in Chapter 2 where we shall see that the present approach lends itself more readily to generalisation when the rather special properties of normal distributions which make the usual linear model the natural one are no longer available.

1.6 An overview of special cases

One of the main purposes of this book is to present a unified account of latent variable models in which existing methods take their place within a single broad framework.

Table 1.3 Classification of latent variable methods.

		Manifest variables	
		Metrical	Categorical
Latent variables	Metrical	Factor analysis	Latent trait analysis
	Categorical	Latent profile analysis	Latent class analysis

This framework can be conveniently set out in tabular form as in Table 1.3. The techniques mentioned there will be defined in later chapters.

It is common to classify the level of measurement of variables as nominal, ordinal, interval or ratio. For our purposes it is convenient to adopt a twofold classification: *metrical* and *categorical*. Metrical variables have realised values in the set of real numbers and may be discrete or continuous. Categorical variables assign individuals to one of a set of categories. They may be unordered or ordered; ordering commonly arises when the categories have been formed by grouping metrical variables. The two-way classification in Table 1.3 shows how the commonly used techniques are related.

It is perfectly feasible to mix types of variables – both manifest and latent. A model including both continuous and categorical *x*s has been given by Moustaki (1996) for continuous *y*s and by Moustaki and Papageorgiou (2004) for categorical *y*s , and these are described in Chapters 6 and 7. When latent variables are of mixed type we obtain what we shall later call *hybrid* models.

1.7 Principal components

We remarked above that the representation

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{1.15}$$

is always possible when $q = p$. This follows from the fact that $\boldsymbol{\Sigma}$ is a symmetric matrix and so can be expressed as

$$\boldsymbol{\Sigma} = \mathbf{M}'\boldsymbol{\Delta}\mathbf{M},$$

where $\boldsymbol{\Delta}$ is a diagonal matrix whose elements are the eigenvalues, $\{\delta_i\}$, of $\boldsymbol{\Sigma}$ and $\mathbf{M}'$ is an orthogonal matrix whose columns are the corresponding eigenvectors of $\boldsymbol{\Sigma}$. Consequently, if we choose

$$\boldsymbol{\Lambda} = \mathbf{M}'\boldsymbol{\Delta}^{1/2} \quad \text{and} \quad \boldsymbol{\Psi} = \mathbf{0},$$

then (1.15) follows. The conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ in (1.11) is now degenerate with all the probability concentrated at the mean given by

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} = \boldsymbol{\mu} + \mathbf{M}'\boldsymbol{\Delta}^{1/2}\mathbf{y}, \tag{1.16}$$

with $\mathbf{x}$ having been expressed exactly as a linear combination of independent standard normal variables. The variables

$$\mathbf{y}^* = \boldsymbol{\Delta}^{1/2}\mathbf{y} = \mathbf{M}(\mathbf{x} - \boldsymbol{\mu}) \tag{1.17}$$

are known as the *principal components*. We shall return to the subject of principal components in Chapter 9 where other approaches will also be presented.

1.8 The historical context

Latent variable models and factor analysis are among the oldest multivariate methods, but their origin and development lie almost entirely outside of mainstream statistics. For this reason topics such as latent class analysis and factor analysis had separate origins and have remained almost totally isolated from one another. This means that they have been regarded as separate fields with virtually no cross-fertilisation. The distinctive feature of the treatment of latent variable models given in this book is that all are presented as special cases of the family described in Section 1.4 and Chapter 2. The statistical approach to such matters, which we have followed, is to begin with a probability model and then use standard statistical ideas to develop the theory. This approach was foreshadowed in Anderson (1959) and taken up again in Fielding (1977), but the first explicit statement, of which we are aware, of the general approach of Section 1.4 was given in Bartholomew (1980) where it was applied to categorical data. The general class of models derived from the exponential family, which underlies the general linear latent variable model of Chapter 2, was proposed in Bartholomew (1984). This, in turn, led to the first edition of the present book (Bartholomew 1987) which demonstrated that most existing latent variable models, and some new ones, could be subsumed under the general framework. The practical potential of this idea has been worked out for an even wider field of applications in Skrondal and Rabe-Hesketh (2004).

All of this means that the extensive literature, extending over almost a century, presents a disjointed picture both notationally and conceptually. It may, therefore, help the reader if we identify some of the key elements in their historical context. This task is made the more necessary by the fact that much of the development has been the work of psychologists and sociologists whose technical language often appears unfamiliar or even meaningless to statisticians. (A good example of this is provided by Maraun (1996) and the ensuing discussion.)

The origins of latent variable modelling can be traced back to the early part of the last century in psychology, notably in the study of human abilities. The origins of factor analysis, in particular, are usually attributed to Spearman (1904). However,

the key idea was already present in Galton (1888) as the following quotation, drawn to our attention by John Aldrich of the University of Southampton, shows:

> Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. . . . It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body.

Whether or not Spearman knew of this must remain a matter for speculation, but an account of Spearman's innovative role in the development of the subject can be found in Bartholomew (1995b). Spearman was concerned with the fact that people, especially children, who performed well in one test of mental ability also tended to do well in others. This led to the idea that all individuals' scores were manifestations of some underlying general ability which might be called general ability or g. However, the scores on different items were certainly not perfectly correlated, and this was explained by invoking factors specific to each item to account for the variation in performance from one item to another. The common and specific factors were thus supposed to combine in some way to produce the actual performance. This idea was expressed in what was originally known as Spearman's two-factor model. He supposed that any test score could be expressed as the sum of two parts – or factors as he would have called them – thus:

$$\text{Score} = \text{common factor} + \text{specific factor}. \tag{1.18}$$

The common factor was named g by Spearman and it appears that he regarded it as a characteristic of the brain, though he deliberately did not call it intelligence in order to avoid the contemporary connotations of that term. If (1.18) were true it could be deduced that the correlation matrix would have a particularly simple form which can be characterised in a variety of ways. One of the most useful was based on the fact that the *tetrad differences* were all zero. If a correlation matrix has typical element ρ_{ij}, the tetrad differences are given by

$$\rho_{ij}\rho_{i+h,j+k} - \rho_{i,j+k}\rho_{i+h,j} = 0,$$

for all integer values of i, j, h, k within the bounds of the matrix.

It was noted quite early by Thomson (1916) that although the two-factor theory was a sufficient condition for the tetrad differences to be zero, it was not necessary. Thomson proposed what has become known as his *bonds model* as an alternative, and we shall return to this in Chapter 9.

At the time the two-factor theory was proposed the idea of a probability model was not current among either statisticians or psychologists. However, to clothe Spearman's idea in modern dress we must translate (1.18) into a statement about the distribution

of random variables. It was probably Bartlett who first recognised that the 'specific factor' of (1.18) had all the characteristics of an error term as used in statistics. From there it is a short step to write (1.18) in the form

$$x_i = \mu_i + \lambda_i y + e_i, \tag{1.19}$$

where x_i is a random variable representing the score on test i, y the common factor and e_i the factor specific to item i. (The notation has been chosen to conform to that used in the early part of this chapter.) If y and the es are mutually independent the correlation coefficient of x_i and x_j will have the form

$$\rho_{ij} = a_i a_j, \tag{1.20}$$

which is the zero tetrad condition in another guise. A a rough test of the plausibility of the model can be made by inspection of the correlation matrix. It follows from (1.20) that

$$\rho_{ih}/\rho_{ij} = a_h/a_j. \tag{1.21}$$

If we form the ratios on the left-hand side for any pair of columns of the matrix they should be constant across rows. Another way of exhibiting the pattern in the matrix is by observing that if $\rho_{ij} \neq 0$ then

$$\rho_{ih}\rho_{hj}/\rho_{ij} = a_h^2, \tag{1.22}$$

for all pairs i and j. Since

$$a_h^2 = \lambda_h^2 \operatorname{var}(y)/\{\lambda_h^2 \operatorname{var}(y) + \operatorname{var}(e_h)\},$$

it is thus possible to estimate the ratio $\lambda_h^2 \operatorname{var}(y)/\operatorname{var}(e_h)$ and so determine the relative contributions of the general and specific factors to the observed score.

If the correlation matrix fails to exhibit the simple pattern required by (1.20) it is natural to add further general factors to the model in the hope of reproducing the observed pattern more exactly.

Within psychology there was a new impetus in the 1930s from Thurstone and his associates in Chicago. He advocated the desirability of obtaining solutions possessing what he called 'simple structure', which meant replacing Spearman's single general factor by a cluster of distinct but possibly correlated factors representing different abilities. The common or general factor of earlier work was thus seen as a sort of average of the 'real' factors and thus of little substantive meaning. Much of the work in this tradition was published in the then new journal *Psychometrika*, the fiftieth anniversary of which was the occasion of a historical review by Mulaik (1986). Although this was primarily concerned with the subject as it has developed in the pages of that journal there are indications of what was happening on a broader front.

Thurstone's writings are now mainly of historical interest, but his book *Multiple Factor Analysis* (Thurstone 1947) is still worth reading for its insight into the essential character and purpose of factor analysis – something which easily becomes overlaid by technical details. The same is true of Thomson's *The Factorial Analysis of Human Ability*. The first edition of this book appeared in 1939 and the fifth and final edition in 1951 with a reprint in 1956; this anticipates many topics which have occupied post-war theorists.

A second fiftieth anniversary, this time of the *British Journal of Mathematical and Statistical Psychology* in 1997, was also marked by a special issue which contains much relevant historical material. Within the psychological tradition there has been a steady stream of texts at all levels, including Harman (1976), Mulaik (1972) and McDonald (1985). Mulaik gives a comprehensive statement of the theory as it then existed, interspersed with illuminating historical material. It is still a useful source of many of the basic algebraic derivations. The book by Harman, which ran to three editions, lays considerable emphasis on statistical methods and computational matters but with less mathematical apparatus than Mulaik or McDonald. However, Mulaik (2009a) is a second edition of the earlier book and brings the subject up to date from the psychological perspective.

A third anniversary occurred in 2004, this time marking 100 years since the publication of Spearman's path-breaking paper of 1904. This was marked by a conference at the Thurstone Laboratory at the University of North Carolina at Chapel Hill in 2004, the 16 contributions to which were published in Cudeck and MacCallum (2007). These represent a diverse set of perspectives looking forward as well as back.

A major treatment of an entirely different kind can be found in the much older book by Cattell (1978) who is at pains to emphasise the limitations of a purely statistical approach to factor analysis as against its use as part of scientific method. His exposition is made without any explicit use of models. It is worth adding that factor analysis has been widely used, mainly by psychologists, for purposes for which it was not intended. We shall note in Chapter 9 its use instead of cluster analysis and multidimensional scaling in Q-analysis. Hotelling (1957) also drew attention to other problems for which factor analysis had been used inappropriately.

The present book, like the first and second editions (Bartholomew 1987; Bartholomew and Knott 1999), lies at the opposite pole to Cattell in that it gives priority to modelling. A model serves to clarify the conceptual basis of the subject and provides a framework for analysis and interpretation. In a subject which has been criticised as arbitrary and hence too subjective, it is especially necessary to clarify in as rigourous a way as possible what is being assumed and what can be legitimately inferred. As already noted, the general approach used here goes back to Anderson (1959), but his work pre-dated the computer era and the time was not ripe for its exploitation.

The first major book-length treatment of factor analysis within the statistical tradition was that of Lawley and Maxwell (1971). The first edition appeared in 1963 and the second, larger, edition followed in 1971; this remains a valuable source of results on the normal theory factor model – especially in the area of estimation and hypothesis testing. Basilevsky (1994) is a more recent comprehensive theoretical

treatment, though *Factor Analysis* in its title is used with a broader connotation than usual. Factor analysis in the sense used here is largely confined to two or three chapters.

More recently the focus of research has shifted to linear structural relations models. These originated with Jöreskog (1970) and may be regarded as a generalisation of the factor model. They incorporate not only the basic factor model but also linear relationships among the *y*s (or factors). Our general framework can be extended to include such models and an introductory account is given in Chapter 8. A key reference is Bollen (1989), but work in this area is focused on three major software packages, LISREL (Jöreskog and Sörbom 2006), Mplus (Muthén and Muthén 2010) and EQS (Bentler 2008). All these authors have made many fundamental contributions to the whole field of latent variable modelling and a selection of their relevant publications is included in the References.

Statisticians have often preferred principal components analysis to factor analysis and some have seen little value in the latter (see, for example, Chatfield and Collins (1980), Hills (1977) and Seber (1984)). The idea of this technique goes back to Pearson (1901) but it was developed as a multivariate technique by Hotelling (1933). Although, as we shall see, it is quite distinct from factor analysis in that it does not depend on any probability model, the close affinities between the two techniques outlined above mean that they are often confused.

Latent structure analysis has its origins in sociology and concerns models in which the latent variables are categorical. Its development has, until recently, been entirely separate from factor analysis. It originated with Lazarsfeld as a tool for sociological analysis and was expounded in the book by Lazarsfeld and Henry (1968). Here again modern computing facilities have greatly extended the applicability of the methods. A more up-to-date account is in Everitt (1984) and Heinen (1996). Langeheine and Rost (1988) and Rost and Langeheine (1997) contain many examples of more recent work.

A third, distinct, strand which has contributed to the current range of models comes from educational testing. The manifest variables in this case are usually categorical, often indicating whether an individual got an item right or wrong in a test. It is often assumed that there is a single continuous latent variable corresponding to an ability of some kind, and the practical objective is to locate individuals on some suitable scale. The contribution of Birnbaum (1968) was a major step forward in this field, which is now the scene of substantial research activity and some controversy. A good introductory account is given by Hambleton *et al.* (1991). Much of the controversy has centred upon the so-called Rasch model (see, for example, Rasch (1960), Andersen (1980b) or Bartholomew (1996)) which has many appealing statistical properties. However, a feature of much of this work on latent trait models, as they are called, which tends to obscure their connection with the general family set out in Section 1.4, is that the latent traits are not treated as random variables. Instead a parameter is introduced to represent each individual's position on the latent scale. Such a model may be termed a *fixed effects* model by analogy with usage in other branches of statistics. Its use would be appropriate if we were interested only in the abilities of the particular individual in the sample and not in the distribution of ability in any

population from which they might have been drawn. It is not easy to find practical examples where this would be the case, and for this reason such models will receive only passing mention in the rest of the book. However, some of the later work on the Rasch model (for example, Andersen (1973, 1977) and Andersen and Madsen (1977)) treats 'ability' as a latent variable and thus falls within our territory. In Chapters 2 and 4, in particular, we shall have occasion to note the simplifications that result from specialisation to the Rasch model.

Fixed effects models have also been considered in factor analysis by Whittle (1953) and Anderson and Rubin (1956), for example, but have not attracted much practical interest. One of their chief disadvantages is that the number of parameters goes up in proportion to the sample size and this creates problems with the behaviour of maximum likelihood estimators. However, there are circumstances in which such methods are relatively simple and can be made to yield estimates of the item parameters which are virtually the same as those derived from a random effects model. They thus have a certain practical interest but in spite of a voluminous and often polemical literature they are, from our standpoint, outside the mainstream of theoretical development.

There is clearly a need for this whole area to be part of current statistical theory and practice. The combination of a unified theoretical treatment of the whole field with the flexibility and power of current computing facilities offers ample scope for new developments in many directions. Some of the gaps in the theory, which were particularly obvious when the first edition of this book (Bartholomew 1987) appeared, have been filled largely by the advent of more adequate computing facilities. But much more remains to be done, especially in introducing to new audiences the unifying ideas on which this book is based.

1.9 Closely related fields in statistics

Latent variables have a long history in statistics, but their ubiquity has been partly obscured by variations in terminology. Lee and Nelder (2009), for example, mention *random effects*, *latent processes*, *factor*, *missing data*, *unobserved future observations* and *potential outcomes*, among others. Even where there is no specific name, the idea is present. For instance, the latent class model is an example of a finite mixture model. There are many such statistical problems where an observed frequency distribution is thought to be composed of a known or unknown number of samples from populations with different distributions. A general treatment of such problems is given in Titterington *et al.* (1985). Mixture models have also been used as a basis for cluster analysis. In fact any analysis based on a latent variable model which assumes that the latent space is categorical can be regarded as a method of cluster analysis.

The idea may be illustrated further by the example of accident proneness which has received much attention in the statistical literature. If all members of a large population have the same small risk of an accident, which is constant over time, then the number of accidents per unit time will have a Poisson distribution. Often it is found that the actual frequency distribution is over-dispersed. This can be explained

by supposing that the Poisson rate parameter is a random variable. Proneness is a construct which cannot be directly observed and so, in our terminology, is a latent variable. It is usual to assume that the rate has a gamma distribution and this leads to a negative binomial distribution for the number of accidents.

More recently, latent variables have been used under the name *hidden variables* especially in the context of discrete time series where Markov chains are used. This term may have been borrowed from quantum mechanics where it is used to refer to unobserved variables which can be invoked, by some, to explain the otherwise random character of quantum behaviour. If the movement of a system among a finite set of states is described by a Markov chain it may happen that the states of the chain cannot be observed directly, but only via manifest indicators. The aim will be to make inferences about the transition matrix of the underlying chain. An introduction to such models is given in Zucchini and MacDonald (2009). This work extends the scope of latent variable modelling into the domain of time series which should prove a fertile field for new developments.

Another long-standing use of latent variables, though under another name, arises in many applications in econometrics and related fields. Econometric models often allow for what is termed unobserved heterogeneity. Discussion of this will be found in the contributions of Chamberlain, Hoem, Trydman and Singer and Tuma to Heckman and Singer (1985) where further references will be found. Another source is Tuma and Hannan (1984).

It is clear that not all the types of latent variable mentioned have the same scientific status. Missing values, for example, are observable *in principle*. They may even have been observed and then lost or mislaid. With finite mixtures the position is more equivocal. We may know that we have a mixture but do not have the means of identifying the underlying populations. On the other hand, we may be exploring the data to see whether the sample could have arisen from a mixture. The latent variables with which we are concerned in this book might be better described as *hypothetical* because, like intelligence, they are constructed for a particular purpose, which is to provide a parsimonious and meaningful description of the data. However, we shall follow well-established custom by using the term *latent*.