ORTHOGONAL MODELS FOR CROSS-CLASSIFIED OBSERVATIONS

REG BUST

Department of Mathematical Statistics

University of Cape Town

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Mathematical Statistics.

ORTHOGONAL MODELS
FOR CROSS-CLASSIFIED OBSERVATIONS

To Sue

ACKNOWLEDGEMENTS

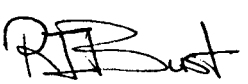
I am particularly indebted to my supervisor, Professor Walter Zucchini, for his wise guidance and assistance, and thank him sincerely.

I would also like to express my gratitude to the members of the Department of Mathematical Statistics, University of Cape Town; in particular to Dr R. Sparks for discussions on the matrix algebra aspects, and to Mrs M. Munitz who made the diagrams. Finally I would like to thank Mrs Tib Cousins, the TEX-pert, for her impressive typesetting achieved despite many lost files and last minute alterations.

This research was funded in part by the CSIR Foundation for Research Development.

DECLARATION

I hereby declare that all the work in this thesis is my own, except where the work of others is referenced.



R. I. BUST

ORTHOGONAL MODELS FOR CROSS-CLASSIFIED OBSERVATIONS

1. INTRODUCTION 1-1

2. PREPARATIONS

1. Model selection 2-1
2. The elements of the problem 2-7
3. Model bases 2-12

3. BASIS MODELS FOR CROSS-CLASSIFIED OBSERVATIONS

1. Basis models 3-1
2. Constructing model bases 3-6
3. The modelling procedure 3-23

4. LINEAR MODELS

1. The model selection procedure 4-1
2. Examples 4-10

5. LOGLINEAR MODELS

1. Simple classifications 5-3
2. Two-way cross-classifications 5-10
3. Multiway cross-classifications 5-27
4. Quasi-hierarchical models 5-33
5. Examples 5-36

6. ROTATION INVARIANCE

1. Linear models 6-1
2. Applications and extensions 6-10
3. Pairwise linear models 6-21
4. Applications and extensions 6-30

APPENDICES

A. THE NEWTON-RAPHSON METHOD A-1

B. PROOFS OF RESULTS IN SECTION 6.1 B-1

C. PROOFS OF RESULTS IN SECTION 6.3 C-1

REFERENCES

CHAPTER 1

INTRODUCTION

This thesis describes methods of constructing models for cross-classified categorical data. In particular we discuss the construction of a class of approximating models and the selection of the most suitable model in the class. Examples of application are used to illustrate the methodology.

The main purpose of the thesis is to demonstrate that it is both possible and advantageous to construct models which are specifically designed for the particular application under investigation. We believe that the methods described here allow the statistician to make good use of any expert knowledge which the client (typically a non-statistician) might possess on the subject to which the data relate.

Presently the participation of the non-statistician in the process of model construction is often confined to the collection of data and to the interpretation of the fitted model, the latter having been supplied by a statistician or by a statistical computer package. The lack of alternatives to the standard parameterisations of models for cross-classified data, in terms of main effects and interactions, reduces the non-statistician's contribution to the statistical analysis to little more than that of specifying the significance level. The methods described here are designed to encourage, and even to demand, the active participation of the client in determining the structure of the models which are to be investigated.

In Chapter 3 we will discuss a class of alternative parameterisations of models for cross-classified observations, namely models in which the parameterisation is completely orthogonal, i.e. each parameter is the coefficient of some "contrast", where the contrasts are orthogonal. The main advantage of such models is that each of the parameters can be treated separately, both for the purposes of model selection and for those of interpretation.

The idea of using orthogonal contrasts is hardly new; it is well-established in the analysis of variance. However, even in that context, the use of orthogonal contrasts

often only occurs at a second stage in the analysis; for example in modelling the relationship between a number of parameters describing a particular main effect which has been shown to be significant. We are proposing that the model be orthogonalised at the start of the analysis.

The process of constructing orthogonal models for a given application is more demanding than simply making use of the standard parameterisations. To obtain meaningful models each contrast must be individually constructed with care. It is here that the expert knowledge of the client can be fruitfully exploited.

For multi-way classifications it is sufficient (and easiest) to separately specify the contrasts for each of the variables which make up the cross-classification. However in some applications more complex parameterisations are appropriate. An example of this type is discussed in which symmetry about the main diagonal of a two-way classification is of interest.

The above discussion relates to the construction of a class of approximating models. We turn now to the question of model selection. There are different approaches to the selection of a model for a given set of observations. One can, for example, select the simplest model which is not (significantly) inconsistent with the observations, i.e. which would not be rejected in a test of the null hypothesis that the data could have been generated by the model. The distribution of the test statistic is therefore derived under the null hypothesis. Although we will not be discussing this approach in the thesis, we observe that no difficulties arise in applying it to the models described here. The distributions of the relevant test statistics can be derived from well-known results with relative ease.

For the purpose of model selection we will adopt the approach described in Linhart and Zucchini (1986a). One begins by specifying a so-called *discrepancy*; a measure of the "difference" or "distance" between models. Selection is then based on the estimated expected discrepancies between approximating models and the "operating" model, the approximating model leading to the smallest of these estimates being selected. The operating model is that which we conceptualise

as actually having produced the data. In this approach to model selection it is at no stage assumed that the operating model might have any simple structure. Thus the object is not to "discover" a simple underlying structure in the operating model, but rather to identify the most appropriate approximating model to fit to the observations.

Two types of models are considered; linear models (Chapter 4) and loglinear models (Chapter 5).

Linear models using orthogonal coefficient matrices are not new and have been considered by Kronmal and Tarter (1968), Ott and Kronmal (1976), Hall (1983), Liang and Krishnaiah (1985a, 1985b) and Diggle and Hall (1986). A particularly convenient discrepancy for the linear case is the sum of squared differences between the corresponding probabilities in the operating and approximating model. In this case the orthogonality property of the parameterisations which we consider leads to a particularly simple model selection algorithm. Each parameter in the saturated model can be considered separately for exclusion or inclusion in the fitted model.

Loglinear models are used more extensively and have a wide literature. Some of the better known references are Goodman (1970), Bishop *et al* (1975), Fienberg (1977), and Plackett (1974). (The parameterisations used by these authors are different from those considered here, and the relationship between the two parameterisations is investigated.) Loglinear models are used more than linear models because with loglinear models:

(i) one can guarantee that the fitted probabilities will lie in the range [0,1];

(ii) one is working with the ratios of probabilities rather than with their differences;

(iii) one can model various forms of independence between groups of variables in a cross-classification.

The Kullback-Leibler discrepancy function (which is intimately linked to the method of maximum likelihood) is a convenient discrepancy for these models. Unfortunately the selection procedure is not so easy in this case. Firstly the expected

discrepancy, which needs to be estimated in order for us to carry out the selection, is difficult to derive. This problem can be circumvented by making use of the cross-validation methods described in Linhart and Zucchini (1986b). The second, and more severe, problem is that the optimal, as well as the estimated, parameter values no longer remain unaffected by the inclusion or exclusion of other parameters into or from the model. Thus it is not possible to simply fit the saturated model and determine which parameters should be included or excluded; one must fit each approximating model which is to be compared. The computational cost of carrying out an exhaustive search for the optimal model increases rapidly as a function of the number of cells in the cross-classification and heuristic methods have to be applied.

The computational cost can be reduced by limiting the search to a subset of models, preferably one in which the models are easy to interpret. Hierarchical models play a special role in this respect, and are well-documented in the literature. However hierarchical models, while being especially convenient when one applies the standard parameterisation, are somewhat restrictive when one uses a completely orthogonal parameterisation. The construction of the orthogonal models is such that it is possible to consider a somewhat larger class than that of hierarchical models, in which all of the models still have clear interpretations. These are described in Section 5.4.

The final chapter deals exclusively with variables whose categories have a cyclical or circular ordering (such as the months of the year). What is required of a modelling procedure in this situation is that the fitted models it produces, for any two rotations of the categories, should be essentially the same, differing only by the corresponding rotation. This property is referred to as *rotation invariance*. For one-way classifications it is shown that rotation invariance can be guaranteed if and only if the number of categories is a power of two. The form of these models is given explicitly in Section 6.1. The extension to multi-way tables in which some or all of the variables are cyclical is given in Section 6.2.

The remainder of the chapter is concerned with cyclical variables which do not

necessarily have 2^m categories. It is shown that, by using a restricted selection procedure, rotation invariance can be achieved for variables with any number of categories. The restriction is that one must exclude or include the parameters in pairs rather than individually. Fourier bases have been applied in this way for decades. More recently and in the context of model selection this type of parameter "pairing" has been used in conjunction with Fourier bases for modelling cyclical variables by Kronmal and Tarter (1968) as well as Linhart and Zucchini (1986a) who are aware that the procedure *"will give the same model no matter which point is taken first"*. We will establish that Fourier bases are effectively the *only* basis with this property; a fact which may not have been previously known.

CHAPTER 2
PREPARATIONS

This chapter contains three sections:

In 2.1 the general approach to statistical modelling which will be used in the thesis will be outlined.

Section 2.2 contains definitions and a brief discussion of those elements of the theory of cross-classified observations which are relevant to the thesis.

In 2.3 we give an outline of a method to construct orthogonal approximating families of models for cross-classified observations.

2.1 MODEL SELECTION

The statistical modelling of a data set begins with what Linhart and Zucchini (1986a) call the *operating model* – the probability model used to conceptualise the process by which the data were generated. Typically the operating model cannot be completely specified, although something of its general form can be ascertained from both the nature of the data and the way in which they were collected.

The aim is to construct a fully specified fitted model which ideally is "close" to the operating model. A fitted model is generally a member of a family of probability models which is indexed by a vector of parameters, say θ, which belong to some parameter space, say Θ. Such a family is called an *approximating family* and written as

$$\mathbf{M} = \{M(\theta) : \theta \in \Theta\}.$$

The fitted model, $M(\hat{\theta})$, is obtained by estimating θ by some estimate $\hat{\theta}$. An approximating family may have the operating model as one of its members or it may only contain models which are simpler than the operating model. Simple models, that is models having a small number of parameters, enjoy certain advantages over models with a large number of parameters. They are generally easier to comprehend and to interpret, and secondly they are less subject to sampling variations than

more complex models. On the other hand simple models are less flexible and may "smooth out" real and intepretable features in the data. It is therefore important to have methods for determining the degree of complexity which is appropriate for the data set under consideration, i.e. methods to select approximating families and models.

In this thesis we will make use of the approach to model selection described in Linhart and Zucchini (1986a). This approach allows for the selection of a fitted model from a *class* of approximating families. The key feature of this methodology is that it is not (necessarily) assumed that the operating model is a member of any of the approximating families under consideration.

Selection is based on a so-called *discrepancy functional*, $\Delta(\cdot,\cdot)$, which is a real-valued non-negative function whose two arguments are probability models. The discrepancy function is in general not symmetric and the probability model corresponding to the second argument is to be considered as an approximation to the first argument, where the extent of the "lack of fit" is given by the value of the discrepancy function. The discrepancy function can be thought of as being to approximation families what the loss function is to estimators.

No specific discrepancy function is prescribed. Linhart and Zucchini (1986a) do list and discuss some of the more widely used discrepancy functions. On the question of the choice of discrepancy function they have this to say:

> *" Whatever strategy is employed to select and fit a model, there will be as a rule a number of aspects in which the operating model and the model which is ultimately fitted differ. Each aspect of the "lack of fit" can be measured by some discrepancy (function) and the relative importance of the different discrepancy (functions) will differ according to the purpose of the envisaged statistical analysis. Consequently, it is proposed that the user should decide which discrepancy (function) is to be minimised."*

Essentially the discrepancy function is constructed to focus on those aspects of the fit which are considered to be important in the context of the particular data

set and the envisaged application of the final fitted model. Of course, mathematical tractability also affects the choice of discrepancy functions.

Suppose then that a discrepancy function has been chosen. Let X denote the data and consider an operating model F and a fitted model $G(\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(X)$ are estimated parameters. For simplicity we identify F with the distribution function of the associated random variable. The discrepancy between the fitted model and the operating model is then given by $\Delta(F, G(\hat{\theta}))$. Nãively one might compare fitted models by comparing their observed discrepancy values. However these observed values are functions of the data and will vary from sample to sample. A compromise is to use the average discrepancy over all possible samples, i.e. the expected discrepancy

$$\int \Delta(F, G(\hat{\theta}))dF.$$

The optimal fitted model from a class of fitted models is then defined to be that which minimises the expected discrepancy.

In order to arrive at a class of fitted models we begin with a class of approximating families, say $\mathbf{S}$, a typical member of which will be denoted by $\mathbf{M}$. The minimum discrepancy parameter for the family

$$\mathbf{M} = \{M(\theta) \ : \ \theta \in \Theta\}$$

is defined by

$$\theta^0 = \arg\{\min_{\theta \in \Theta} \Delta(F, M(\theta))\}.$$

This minimum discrepancy parameter for the family $\mathbf{M}$, which is a function of the unknown operating model, is then estimated, by say $\hat{\theta}(X)$, so that $M(\hat{\theta}) \in \mathbf{M}$. In principle any method of estimation may be used. In fact for a single approximating family two or more different estimation procedures can be used and compared by comparing the expected discrepancies of the associated fitted models. However one estimation procedure presents itself as the natural one to use. It consists of estimating θ^0 by

$$\hat{\theta} = \arg\{\min_{\theta \in \Theta} \Delta(F_n, M(\theta))\}$$

where F_n is the empirical distribution function.

In this thesis we will consider only *minimum empirical discrepancy estimators.* In effect this reduces the problem of model selection to that of selecting an approximating family, because the latter, together with the discrepancy function, determine $\hat{\theta}$ and hence the fitted model.

In summary, we begin with a class of approximating families $\mathbf{S}$. From each approximating family, $\mathbf{M} \in \mathbf{S}$, in the class one fitted model $M(\hat{\theta}) \in \mathbf{M}$ is obtained. The optimal fitted model is that with the smallest expected discrepancy

$$\int \Delta(F, M(\hat{\theta}))dF$$

of all the fitted models. Clearly, since F is unknown, we are not in a position to determine which fitted model is optimal. What we can do is to estimate the expected discrepancy for different fitted models and thereby determine which model is *estimated* to be optimal. This will be discussed in more detail below.

Interpretation. A useful interpretaton of the selection procedure which has been outlined above is now given. (The interpretation given here differs slightly from that given by Linhart and Zucchini (1986a). They deal with discrepancy values while we prefer to deal with expected discrepancies as it is the expected discrepancy which is used to compare fitted models.)

The interpretation is based on the decomposition of the expected discrepancy into two components. The first of these is called the *discrepancy due to approximation*, $\Delta(F, M(\theta^0))$, where θ^0 is such that $\Delta(F, M(\theta^0)) \leq \Delta(F, M(\theta))$ for all $\theta \in \Theta$. Thus, in terms of the chosen discrepancy, $M(\theta^0)$ is the best approximating model in the family $\mathbf{M}$. In practice we are not in a position to determine θ^0 since this would presuppose that we know F, or at least $\Delta(F, M(\theta))$, but we can estimate θ^0, and thus $M(\theta^0)$, using the data. The fitted model, $M(\hat{\theta})$, will differ from $M(\theta^0)$ in general. The difference can be quantified in terms of the discrepancy, i.e. by $\Delta(M(\theta^0), M(\hat{\theta}))$ which is a random variable and which is called

the *discrepancy due to estimation.* The expected discrepancy due to ~~approximation~~ estimation

$$\int \Delta(M(\theta^0), M(\hat{\theta}))dF$$

thus gives the average "lack-of-fit" arising from sampling variation (rather than the limitations of the models in the family to match the operating model).

The discrepancy due to approximation quantifies the potential accuracy of the approximation and can be thought of as a generalisation of the bias of an estimator. The expected discrepancy due to estimation on the other hand, measures the variability introduced by having to estimate the parameters and is a generalisation of the concept of the variance of an estimator.

For a number of discrepancy functions the expected discrepancy can be decomposed into the sum of these two components (plus possibly one other inessential term). For all discrepancy functions the expected discrepancy is some function of these two components which is such that the two components act in opposition to each other in the sense that decreasing either one tends to increase the other. In finding the fitted model which minimises the expected discrepancy one is achieving an optimal compromise between the two components.

Large approximating families with many parameters are likely to contain a model close to the operating model which means that the discrepancy due to approximation will be small. In fact if the approximating family contains the operating model then the ~~expected~~ discrepancy due to approximation is zero. In general, however, the more parameters there are to estimate, the less reliably they can be estimated so that the expected discrepancy due to estimation will tend to be large. Conversely approximating families having few parameters will tend to have a large discrepancy due to approximation and a smaller discrepancy due to estimation.

Estimating expected discrepancies. In practice one cannot use the expected discrepancy to make comparisons as it is not fully known. Consequently the expected discrepancy itself has to be estimated. (The problem is akin to that which leads to parameter estimators which minimise an estimate of the expected loss

rather than the loss itself.) Linhart and Zucchini (1986a) admit: "*This estimation problem is the weakest link in the selection procedure based on discrepancies.*" In most cases it is extremely difficult to obtain "good" estimates of an expected discrepancy. No one procedure for estimating the expected discrepancy is prescribed and different approaches are taken for different operating models and different discrepancy functions. However cross-validatory estimation procedures are particulary convenient for many discrepancy functions (Linhart and Zucchini, 1986b). •

2.2 THE ELEMENTS OF THE PROBLEM

The particular field to which the approach outlined in the previous section will be applied, ~~to~~ is that of cross-classified observations. We consider individuals in a population each of which can be described by a number of attributes or variables. Attention is restricted to categorical variables which have a finite number of categories. The totality of different descriptions is called the *cross-classification.* The number of variables involved determines, in part, the way that the cross-classification is presented. For a single variable a simple list of categories, usually called a *classification*, will suffice. For two variables the usual way to present the cross-classification is as a rectangular table with columns corresponding to the categories of one of the variables and rows to the categories of the other variable. Three variables leads to layers of two-way tables, and so on.

The observations consist of a table of counts. Each entry in the table gives the number of individuals in the sample having the same description, i.e. which can be classified in the same way.

A cross-classification involving K variables is called a $K-$way cross-classification or a $K-$way table. If the categories associated with the kth variable $(k = 1,\ldots,K)$ are labelled as $1,2,\ldots,L_k$ then the cells in the cross-classification can be identified by

$$(i_1,i_2,\ldots,i_K) \quad \text{with} \quad i_k \in \{1,2,\ldots,L_k\} \quad \text{for } k = 1,\ldots,K.$$

The resulting cross-classification is sometimes referred to as a $L_1 \times L_2 \times \cdots \times L_K$ table.

Often it is convenient to present a $K-$way cross-classification as a one-way classification. To do this one has to order the cells in the K -way table. A method of ordering the categories which is particularly convenient in our context is lexicographic ordering and will be used throughout the thesis.

Definition. A real $K-$tuple $(i_1,i_2,\ldots,i_K)$ is said to be *lexicographically* less than a real $K-$tuple $(j_1,j_2,\ldots,j_K)$ iff for the smallest u $(1 \leq u \leq K)$ such

that $i_u \neq j_u$ we have $i_u < j_u$. •

Example. Consider a 2×2 cross-classification where the cells are represented as

$$(1,1) \quad (1,2)$$
$$(2,1) \quad (2,2).$$

Arranging these cells into a vector whose elements are lexicographically ordered gives

$$(1,1)$$
$$(1,2)$$
$$(2,1)$$
$$(2,2).$$

•

The device of transforming a multiway cross-classification into a one-way classification is used frequently and often implicitly, in the sequel.

OPERATING MODELS

The operating model is the conceptual model that one uses in planning and thinking about the experiment. In this thesis the form of the operating model depends on two factors:

(i) the sampling scheme used; and

(ii) the way we view each of the variables in the cross-classification.

These are now discussed.

(i) **Sampling schemes.** We will consider two sampling schemes. The first is where one random sample of fixed size is taken from the population, and is called *multinomial sampling.* We define

(1) n_i as the number of sampled units which fall into the ith category,

(2) $n_+ = \sum_{i=1}^{L} n_i$, the total size of the sample,

(3) $p_i = n_i/n_+$, the proportion of sampled units which fall into the ith category.

To apply the second sampling scheme one needs two variables, say X and Y, (each of which may be multivariate) whose classifications will be taken to be $\{x_1,\ldots,x_R\}$ and $\{y_1,\ldots,y_C\}$ respectively. One of the variables, say Y, is used to form a partition of the population into *sub-populations.* The jth sub-population, for $j = 1,\ldots,C$ consists of all those members in the population which can be classified as having $Y = y_j$. From each of these sub-populations an independent random sample is taken. This scheme is called *product-multinomial sampling.* We define

(1) n_{ij} as the number of units in the jth sub-population sample having $X = i$,

(2) $n_{+j} = \sum_{i=1}^{R} n_{ij}$, the size of the sample taken from the jth sub-population,

(3) $p_{i(j)} = n_{ij}/n_{+j}$, the proportion of units in the jth sub-population sample having $X = i$.

(ii) **Response and explanatory variables.** The variables in a cross-classification are often classified as being either *response variables* or *explanatory variables.* The purpose of this distinction is to emphasize that one is sometimes interested in the conditional distribution of a certain set of the variables (the response variables) for given levels of the rest of the variables (the explanatory variables). It is not always clear which variables should be regarded as the response, and which as the explanatory, variables. How the variables should be classified depends largely on the purpose of the analysis. Fortunately however, the model selection procedures discussed here are essentially unaffected by the classification of variables as response or explanatory.

Two types of operating model will be considered. The first being that where all the variables involved in the cross-classification are considered to be response variables and multinomial sampling is employed. In this case we will say that the *operating model is multinomial.* Let X and $\{x_1,\ldots,x_L\}$ denote the (multi-dimensional) variable and its associated (cross-) classification. X can then be considered to be a *random* variable which assumes values in the range $\{x_1,\ldots,x_L\}$. The probability with which a randomly selected member of the population falls into

the ith category is then the probability with which the random variable X assumes the value x_i, which is denoted by π_i for $i = 1, ..., L$. Since each member must fall into one of the categories, the π_i must satisfy $\sum_{i=1}^{L} \pi_i = 1$. The π_i are called the operating model probabilities and are generally unknown. The vector $\underline{\pi}$ is called the *(multinomial) operating model vector*; it completely specifies the operating model.

The second type of operating model makes allowance for both response and explanatory variables. Let X and Y denote the possibly multidimensional response and explanatory variables respectively, whose respective classifications will be taken to be $\{x_1, \ldots, x_R\}$ and $\{y_1, \ldots, y_C\}$. For each y_j, $j = 1, \ldots, C$, introduce a random variable $X_{(j)}$ which assumes values in the range $\{x_1, \ldots, x_R\}$. The probability with which a member of the jth sub-population, selected at random, has an X value of x_i can then be viewed as the probability with which the random variable $X_{(j)}$ assumes the value x_i, and will be denoted by $\pi_{i(j)}$. In this case we will say that the *operating model is product-multinomial.* The operating model probabilities are the $\pi_{i(j)}$ where these satisfy

$$\sum_{i=1}^{R} \pi_{i(j)} = 1 \quad \text{for} \quad j = 1, \ldots, C.$$

Let π denote the $R \times C$ matrix of indexed probabilities

$$\begin{pmatrix} \pi_{1(1)} & \pi_{1(2)} & \cdots & \pi_{1(C)} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{R(1)} & \pi_{R(2)} & \cdots & \pi_{R(C)} \end{pmatrix}$$

whose columns will be denoted by

$$(\underline{\pi}_{(1)}, \underline{\pi}_{(2)} \cdots, \underline{\pi}_{(C)}).$$

The vector

$$\underline{\pi} = \begin{pmatrix} \underline{\pi}_{(1)} \\ \underline{\pi}_{(2)} \\ \vdots \\ \underline{\pi}_{(C)} \end{pmatrix} \tag{1}$$

is called the *(product-multinomial) operating model vector*; it completely specifies the operating model.

Clearly the multinomial operating model is a special case of the product-multinomial in which $C = 1$. •

2.3 MODEL BASES

For the types of operating models considered in this thesis, namely multinomial and product-multinomial, a particularly simple and convenient way of constructing approximating models is to impose linear constraints on the vector of operating probabilities. One way of doing this is to write the vector of operating probabilities in terms of a basis in $\Re^L$, i.e. as a linear combination of independent vectors in $\Re^L$ where the coefficients of the vectors are regarded as the unknown parameters. Approximating families are then obtained by setting some of the parameters equal to zero. By suitable selection of the vectors in such a basis it is possible to construct approximating models which lead to simple model selection methods; and to approximating models which are easy to interpret. In particular it is especially convenient to work with a special kind of orthonormal basis – called a "model basis".

Before defining a model basis we will need to give some notation and (standard) definitions.

(1) A vector in $\Re^L$ will be written as

$$\underline{u} = [u_i]_{i=1,\ldots,L} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_L \end{pmatrix} .$$

(2) The dot product of two vectors $\underline{\phi}_q = [\phi_{iq}]_{i=1,\ldots L}$ and $\underline{\phi}_r = [\phi_{ir}]_{i=1,\ldots,L}$ is defined by

$$\underline{\phi}_q \cdot \underline{\phi}_r = \Sigma_i \phi_{iq} \phi_{ir} = \underline{\phi}_q' \underline{\phi}_r$$

where $\underline{\phi}_q'$ is the transpose of the vector $\underline{\phi}_q$.

(3) A set of L linearly independent vectors in $\Re^L$, $(\underline{\phi}_1, \ldots, \underline{\phi}_L)$, forms a basis for $\Re^L$ in that any vector in $\Re^L$ can be written as a linear combination of the $\{\underline{\phi}_q\}$.

Definition. A *model basis* for $\Re^L$ say $(\underline{\phi}_1, \ldots, \underline{\phi}_L)$ is defined as a basis for $\Re^L$ which is such that

(1) the basis is orthogonal, i.e. for all $\underline{\phi}_q$ and $\underline{\phi}_r$ in the basis

$$\underline{\phi}_q \cdot \underline{\phi}_r = 0 \quad \text{if} \quad q \neq r$$

(2) the vectors in the basis are normalised, i.e.

$$\underline{\phi}_q \cdot \underline{\phi}_q = 1 \quad \text{for} \quad q = 1, \ldots, L$$

(3) $\underline{\phi}_1 = \frac{1}{\sqrt{L}} \underline{1}_L$ where $\underline{1}_L = [1]_{i=1,\ldots,L}$. •

A simple standard result which is central to the modelling procedure is: given a vector in $\Re^L$, for example, $[g(\pi_i)]_{i=1,\ldots,L}$, where g is some real-valued function, and given an orthonormal basis for $\Re^L$, say $(\underline{\phi}_1, \ldots, \underline{\phi}_L)$, one may write

$$[g(\pi_i)]_{i=1,\ldots,L} = \sum_{q=1}^{L} \underline{\phi}_q \theta_q,$$

i.e.

$$g(\pi_i) = \sum_{q=1}^{L} \phi_{iq} \theta_q \quad \text{for} \quad i = 1, \ldots, L \tag{1}$$

where the θ_q are unique and, in fact,

$$\theta_q = \sum_{i} \phi_{iq} g(\pi_i) \quad \text{for} \quad q = 1, \ldots, L. \tag{2}$$

Example. Let $[g(\pi_i)]_{i=1,2,3}$ be given by

$$\begin{pmatrix} .5 \\ .2 \\ .3 \end{pmatrix}$$

and suppose that we have an orthonormal basis

$$(\underline{\phi}_1, \underline{\phi}_2, \underline{\phi}_3) = \begin{pmatrix} 1/\sqrt{3} & 2/\sqrt{6} & 0 \\ 1/\sqrt{3} & -1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{6} & -1/\sqrt{2} \end{pmatrix}.$$

Expressing $[g(\pi_i)]_{i=1,2,3}$ in terms of the basis gives

$$\begin{pmatrix} .5 \\ .2 \\ .3 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix} + \frac{.5}{\sqrt{6}} \begin{pmatrix} 2/\sqrt{6} \\ -1/\sqrt{6} \\ -1/\sqrt{6} \end{pmatrix} - \frac{.1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

One can think of this expansion in the following way. One begins with

$$\theta_1 \underline{\phi}_1 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$$

in which all of the elements are equal. Thereafter including terms involving contrast vectors will lead to differentiation among the elements. In fact

$$\theta_1 \underline{\phi}_1 + \theta_2 \underline{\phi}_2 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} + \frac{.5}{6} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} .5 \\ .25 \\ .25 \end{pmatrix}$$

while

$$\theta_1 \underline{\phi}_1 + \theta_3 \underline{\phi}_3 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} - .05 \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2/6 \\ 1.7/6 \\ 2.3/6 \end{pmatrix}.$$

SOME MODEL BASES

Two standard types of orthogonal bases which can be transformed into model bases and which will be used later, namely orthogonal polynomial bases and Hadamard bases, are now defined. These are by no means the only model bases that will be used. Other model bases will be introduced in conjunction with specific variables that appear in examples of applications c.f. Section 3.2.

Orthogonal polynomial bases. Orthogonal polynomial bases have been used in factorial designs (Raktoe, Hedayat, Federer, 1981). They can be used in connection with variables which assume quantifiable values. A typical example would be the different dosage levels at which a drug was administered to various experimental units.

Let X be a categorical random variable which can assume the values $x_1, x_2, \ldots, x_L$ where these values are quantifiable. The *orthonormal polynomial*

basis for X can be defined as the matrix produced by the Gram-Schmidt orthonormalisation procedure, when applied to the matrix

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{L-1} \\ 1 & x_2 & \dots & x_2^{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_L & \dots & x_L^{L-1} \end{pmatrix}.$$

It can be shown that an orthonormal polynomial basis is a model basis. An important result is that the orthogonal polynomial bases are invariant with respect to location-scale transformations of the x_i ; in the sense that the orthonormal polynomial bases obtained when the input values are x_i or $ax_i + b$ with $a, b \in \Re$ are identical. This result means that in all cases when the x_i are equispaced (i.e. $|x_i - x_j|$ is constant for all $i, j;\ i \neq j$) ; the same orthonormal polynomial basis is appropriate (and can be obtained by applying the Gram-Schmidt process to a variable whose values are coded as $1, \dots, L$) .

Hadamard bases. Another type of model basis which we will find useful are the so-called Hadamard bases which are used extensively in 2^n factorial design experiments.

These bases can be introduced as follows: Consider choosing a model basis for a binary random variable. By the definition of a model basis it follows that the basis must be of the form

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & cx \\ \frac{1}{\sqrt{2}} & c(-x) \end{pmatrix}$$

where $x \in \Re$ and c is a normalising factor chosen such that the second column vector has length 1. It is convenient to take $x = 1$ so that the basis becomes

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We call this the (normalised) Hadamard matrix of dimension 2, and denote it by H_2 .

Hadamard bases of higher dimension are constructed from smaller ones by taking so-called "left Kroneker products".

The *left Kroneker product* or simply the *Kroneker product* of two matrices

$$A = [a_{ij}]_{i=1,\ldots m;\ j=1,\ldots,n} \quad \text{and}$$

$$B = [b_{ij}]_{i=1,\ldots r;\ j=1,\ldots,s}$$

is defined as the $mr \times ns$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ a_{21}B & a_{22}B & \ldots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \ldots & a_{mn}B \end{pmatrix}.$$

With this definition it can now be stated that the (normalised) Hadamard matrices of dimension 2^n are defined inductively by

$$H_{2^n} = H_{2^{n-1}} \otimes H_2.$$

The fact that the H_{2^n} are model bases follows from the fact that the Kroneker product of two model bases is again a model basis.

MULTIVARIATE CONSIDERATIONS

Given a multiway table with a total of say L cells one can transform the table of probabilities into a single vector and express this vector in terms of a model basis for $\Re^L$. However in order to emphasise the relationship between the variables in the cross-classification it is often preferable to maintain the multivariate notation. We thus wish to determine the multivariate analogs of (1) and (2).

Consider a two-way cross-classification with R rows and C columns and let both the row and column variable each have its own basis, say $\Psi = [\psi_{ir}]_{i=1,\ldots,R;\ r=1,\ldots,R}$ and $\Omega = [\omega_{jc}]_{j=1,\ldots,C;\ c=1,\ldots,C}$ respectively. Then the elements of $[g(\pi_{ij})]_{i=1,\ldots,R;\ c=1,\ldots,C}$ can be expressed as

$$\left.\begin{aligned} g(\pi_{ij}) &= \sum_{r=1}^{R}\sum_{c=1}^{C} \psi_{ir}\omega_{jc}\theta_{rc} \quad \text{for all } i,j \\ \text{where} \quad \theta_{rc} &= \sum_{i=1}^{R}\sum_{j=1}^{C} \psi_{ir}\omega_{jc}g(\pi_{ij}) \quad \text{for all } r,c \end{aligned}\right\} \qquad (3)$$

In order to see how (3) is arrived at

(a) put $\Phi = \Psi \otimes \Omega$ and label the columns of Φ as $\underline{\phi}_{11}, \ldots, \underline{\phi}_{1C}, \underline{\phi}_{21}, \ldots, \underline{\phi}_{RC}$ so that

$$\underline{\phi}_{rc} = \begin{pmatrix} \psi_{1r} \\ \vdots \\ \psi_{Rr} \end{pmatrix} \otimes \begin{pmatrix} \omega_{1c} \\ \vdots \\ \omega_{Cc} \end{pmatrix} = \begin{pmatrix} \psi_{1r}\omega_{1c} \\ \vdots \\ \psi_{1r}\omega_{Cc} \\ \vdots \\ \psi_{Rr}\omega_{1c} \\ \vdots \\ \psi_{Rr}\omega_{Cc} \end{pmatrix}, \tag{4}$$

(b) arrange the matrix of π_{ij} into a single vector where the indices are ordered lexicographically, i.e.

$$\underline{\pi} = \begin{pmatrix} \pi_{11} \\ \vdots \\ \pi_{1C} \\ \vdots \\ \pi_{R1} \\ \vdots \\ \pi_{RC} \end{pmatrix} \tag{5}$$

and label the elements of $\underline{\pi}$ as π_a for $a = 1, \ldots, RC$.

Then $[g(\pi_a)]_{a=1,\ldots,RC}$ is an element of $\Re^{RC}$ for which Φ is a model basis. Thus the analogs of (1) and (2) are

$$\left.\begin{aligned} g(\pi_a) &= \sum_{r=1}^{R}\sum_{i=1}^{C} \theta_{rc}\phi_{a,rc} \\ \text{with} \quad \theta_{rc} &= \underline{\phi}_{rc} \cdot [g(\pi_a)]_{a=1,\ldots,RC} \end{aligned}\right\} \tag{6}$$

By rewriting this in two-subscript notation it follows from (4) and (5), that one can substitute π_{ij} for π_a at the same time that $\psi_{ir}\omega_{jc}$ is substituted for $\phi_{a,rc}$. Making these substitutions into (6) yields (3).

Note that all of the foregoing applies equally to multinomial and product-multinomial probabilities and one can replace π_{ij} with $\pi_{i(j)}$ in (6).

Expression (3) can be rewritten in the following form in which the role of the

individual parameters is more transparent:

$$g(\pi_{ij}) = \theta_{11} + \sum_{r=2}^{R} \psi_{ir}\theta_{r1} + \sum_{c=2}^{C} \omega_{jc}\theta_{1c} + \sum_{r=2}^{R}\sum_{c=2}^{C} \psi_{ir}\omega_{jc}\theta_{rc}$$

with

$$\theta_{11} = \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C} g(\pi_{ij})$$

$$\theta_{r1} = \sum_{i=1}^{R} \psi_{ir}\left(\sum_{j=1}^{C} g(\pi_{ij})\right) \quad \text{for } r = 2,\ldots,R$$

$$\theta_{1c} = \sum_{j=1}^{C} \omega_{jc}\left(\sum_{i=1}^{R} g(\pi_{ij})\right) \quad \text{for } c = 2,\ldots,C$$

$$\theta_{rc} = \sum_{i=1}^{R}\sum_{j=1}^{C} \psi_{ir}\omega_{jc}g(\pi_{ij}) \quad \text{for } r = 2,\ldots,R;\ c = 2,\ldots,C.$$

It can be seen that:

(1) θ_{11} is the average

(2) each θ_{r1} (for $r = 2,\ldots,R$) is a contrast of the row marginals $\{\Sigma_j g(\pi_{ij})\}_{i=1,\ldots,R}$ using the contrast vector $\underline{\psi}_r$

(3) similarly each θ_{1c} is a contrast involving $\underline{\omega}_c$ of the column marginals $\{\Sigma_i g(\pi_{ij})\}_{j=1,\ldots,C}$

(4) the remaining θ_{rc} involve contrasts of the individual cells across both rows and columns.

An example will further illustrate these roles.

Example. Let g be the identity and let $[\pi_{i(j)}]_{i=1,2,3,\ j=1,2}$ be given by

.5	.5	.5
.3	.25	.275
.2	.25	.225
1	1	

and let the model bases used for the row and column variables be

$$(\underline{\phi}_1, \underline{\phi}_2, \underline{\phi}_3) = \begin{pmatrix} 1/\sqrt{3} & 2/\sqrt{6} & 0 \\ 1/\sqrt{3} & -1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{6} & -1/\sqrt{2} \end{pmatrix}$$

and

$$(\underline{\omega}_1, \underline{\omega}_2) = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

respectively. We may then write

$$\begin{pmatrix} .5 & .5 \\ .3 & .25 \\ .2 & .25 \end{pmatrix} = \theta_{11} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} + \left[\theta_{21} \begin{pmatrix} 2/\sqrt{6} & 2/\sqrt{6} \\ -1/\sqrt{6} & -1/\sqrt{6} \\ -1/\sqrt{6} & -1/\sqrt{6} \end{pmatrix} + \theta_{31} \begin{pmatrix} 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \right]$$

$$+ \left[\theta_{12} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \right]$$

$$+ \left[\theta_{22} \begin{pmatrix} 2/\sqrt{12} & -2/\sqrt{12} \\ -1/\sqrt{12} & 1/\sqrt{12} \\ -1/\sqrt{12} & 1/\sqrt{12} \end{pmatrix} + \theta_{32} \begin{pmatrix} 0 & 0 \\ 1/\sqrt{4} & -1/\sqrt{4} \\ -1/\sqrt{4} & 1/\sqrt{4} \end{pmatrix} \right]$$

where

$$\theta_{11} = \frac{1}{3}$$

$$\theta_{21} = \frac{.5}{\sqrt{6}} \quad , \qquad \theta_{31} = \frac{.05}{\sqrt{2}}$$

$$\theta_{12} = 0$$

$$\theta_{22} = 0 \quad , \qquad \theta_{32} = \frac{.1}{\sqrt{4}}.$$

Remarks. (1) In the coefficient matrices of the parameters θ_{r1} for $r = 1,2,3$ both columns are identical. Furthermore

$$\theta_{11} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} + \theta_{21} \begin{pmatrix} 2/\sqrt{6} & 2/\sqrt{6} \\ -1/\sqrt{6} & -1/\sqrt{6} \\ -1/\sqrt{6} & -1/\sqrt{6} \end{pmatrix} + \theta_{31} \begin{pmatrix} 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} .5 & .5 \\ .275 & .275 \\ .225 & .225 \end{pmatrix}$$

where each column is identical and matches the original row marginals exactly.

(2) The rows in the coefficient matrices of the parameters θ_{1c}, for $c = 1,2$, are identical and

$$\theta_{11} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} + \theta_{12} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 1/3 \\ 1/3 & 1/3 \end{pmatrix}.$$

(3) The coefficient matrix of θ_{22} can be viewed in two ways; either as having columns

$$\underline{\phi}_2 = 1/\sqrt{6}\begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$$

contrasted according to $\underline{\omega}_2 = 1/\sqrt{2}(1,-1)'$, or as having rows $\underline{\omega}_2$ contrasted according to $\underline{\phi}_2$

(4) Remark (3) also applies to the coefficient matrix of θ_{32} but with $\underline{\phi}_2$ now replaced with $\underline{\phi}_3$. •

Having seen the extension from the univariate to the bivariate situation, the extension to further dimensions is straight-forward. In fact, if we have a $K-$way cross-classification with K model bases $\Phi^{(1)},\ldots,\Phi^{(K)}$ then we may write

$$\pi_{i_1\cdots i_K} = \sum_{a_1}\cdots\sum_{a_K}\theta_{a_1\cdots a_K}\phi^{(1)}_{i_1a_1}\cdots\phi^{(K)}_{i_Ka_K}$$

with

$$\theta_{a_1\cdots a_K} = \sum_{i_1}\cdots\sum_{i_K}\phi^{(1)}_{i_1a_1}\cdots\phi^{(K)}_{i_Ka_K}\pi_{i_1\cdots i_K}. \quad \bullet$$

CHAPTER 3

BASIS MODELS FOR CROSS-CLASSIFIED OBSERVATIONS

In this chapter we introduce two classes of approximating families for use in modelling cross-classified observations. The models in these approximating families express some function of the modelled cell probabilities as a linear combination of parameters. The coefficient matrices in the linear combinations are model bases, hence the name *basis models.*

The construction of the approximating families is given in Section 1. In Section 2 attention is given to the construction of the model bases for particular data sets, while Section 3 considers the model selection problem of finding the optimal fitted model from within a given class of approximating families.

3.1 BASIS MODELS

Consider a multinomial operating model's vector of probabilities $\underline{\pi} = [\pi_i]_{i=1,\ldots,L}$. A class of approximating families can be constructed by specifying a model basis for $\Re^L$ and an invertible real-valued function g, called the *link function.* Discussion on the choice of basis and of the link function is delayed until the classes of approximating families have been fully introduced. It was shown in Chapter 2 that given $[g(\pi_i)]_{i=1,\ldots,L} \in \Re^L$ and a model basis $(\underline{\phi}_1, \ldots, \underline{\phi}_L)$ for $\Re^L$ we can write

$$g(\pi_i) = \sum_{q=1}^{L} \phi_{iq}\theta_q \quad \text{for } i = 1, \ldots, L.$$

The next step is to consider excluding some of the parameters from the expansion to arrive at models $[M_i(\theta)]_{i=1,\ldots,L}$ for $[\pi_i]_{i=1,\ldots,L}$ where

$$g(M_i(\theta)) = \sum_{q \in Q} \phi_{iq}\theta_q \quad \text{for } i = 1, \ldots, L$$

where $Q \subseteq \{1, \ldots, L\}$ and the parameters are subject to the constraint $\Sigma_i M_i(\theta) = 1$. Each Q determines an approximating family which can be defined by

$$\mathbf{M}(Q) = \left\{ [M_i(\theta)]_{i=1,\ldots,L} \; : \; g(M_i(\theta)) = \sum_{q \in Q} \phi_{iq}\theta_q \; ; \; \sum_i M_i(\theta) = 1 \right\} \qquad (1)$$

The approximating family $\mathbf{M}(\{1,\ldots,L\})$ is the only approximating family which necessarily contains the operating model. Other families will generally have larger discrepancies due to approximation but smaller expected discrepancies due to estimation. A class of approximating families is obtained by considering various Q. Generally this class is made as large as is possible. However for different link functions one may need to insist that Q contain a particular index, such as 1, to ensure that the constraint $\Sigma_i M_i(\theta) = 1$ can be satisfied. If $\mathbf{S}(L)$ denotes the set through which Q ranges then the class of approximating families is defined by $\{\mathbf{M}(Q) : Q \in \mathbf{S}(L)\}$.

Consider next a two-way $R \times C$ multinomial operating model with cell probabilities $[\pi_{ij}]_{i=1,\ldots,R;\ j=1,\ldots,C}$. While we can regard this case as being formally equivalent to that considered above it is often preferable not to, but rather to emphasise the bivariate nature of the operating model. Suppose then that we have a model basis $[\psi_{ir}]_{i=1,\ldots,R;\ r=1,\ldots,R}$ for the row variable and a separate model basis $[\omega_{jc}]_{j=1,\ldots,C;\ c=1,\ldots C}$ for the column variable. If Q now denotes a subset of $\mathbf{R} \times \mathbf{C} = \{(r,c) : r \in \{1,\ldots,R\}, c \in \{1,\ldots,C\}\}$, then a typical approximating family can be written as

$$\mathbf{M}(Q) = \Big\{ [M_{ij}(\theta)]_{i=1,\ldots,R;\ j=1,\ldots C} : g(M_{ij}(\theta)) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc};\ \sum_i \sum_j M_{ij}(\theta) = 1 \Big\}. \tag{2}$$

A class of approximating families is obtained by varying Q through some collection of subsets of $\mathbf{R} \times \mathbf{C}$.

Consider now an $R \times C$ product-multinomial operating model with indexed probabilities $[\pi_{i(j)}]_{i=1,\ldots,R;\ j=1,\ldots,C}$. Here one *must* use a separate basis for the response and explanatory variables. A typical approximating family is then

$$\mathbf{M}(Q) = \Big\{ [M_{i(j)}(\theta)]_{i=1,\ldots,R;\ j=1,\ldots,C} : g(M_{i(j)}(\theta)) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc};\ \sum_i M_{i(j)}(\theta) = 1 \quad \text{for } j = 1,\ldots,C \Big\} \tag{3}$$

for some $Q \subseteq R \times C$. The C restrictions $\sum_i M_{i(j)}(\theta) = 1$ lead to these classes of approximating families generally containing fewer members than their multinomial counterparts.

We now illustrate the role that the parameters play in the above models. The two examples of Section 2.3 should be borne in mind. Consider initially a model $[M_i(\theta)]_{i=1,\ldots,L}$ where

$$g(M_i(\theta)) = \sum_{q \in Q} \phi_{iq}\theta_q \quad \text{for } i = 1, \ldots, L$$

One way of viewing the model is this:

- the inclusion of the parameter θ_q in the model means adding a contribution to each of the cells, namely $\phi_{iq}\theta_q$ to the ith cell for $i = 1, \ldots, L$,
- in each cell the sum of these contributions gives $g(M_i(\theta))$,
- the modelled probability $M_i(\theta)$ is obtained from $g(M_i(\theta))$.

The role that θ_q plays in this process is effectively determined by $\underline{\phi}_q$, in that the inclusion of θ_q in a model means adding $\theta_q \underline{\phi}_q$ to the cells. Since $\underline{\phi}_1 = \frac{1}{\sqrt{L}}\underline{1}$, while the other $\underline{\phi}_q$ are contrast vectors, it follows that θ_1 plays a different role from the other parameters. In fact θ_1 generally appears in all models, and its role is to make an equal contribution to each of the cells; which we can think of as providing a model – namely the simplest possible model $g(M_i(\theta)) = g(\frac{\theta_1}{\sqrt{L}}) = k$ for all i , for some constant k. On the other hand the introduction of a parameter θ_q with $q \neq 1$ into a model leads to different contributions, some positive and some negative, being made to different cells and results in differentiation between the modelled cell probabilites. It should be pointed out that for any θ_q with $q \neq 1$ the contributions $\{\phi_{iq}\theta_q\}_{i=1,\ldots,L}$ always sum to zero.

Similar considerations apply to two-way tables. Here a typical model is

$$g(M_{ij}(\theta)) = \sum_{(r,c) \in Q} \psi_{ir}\omega_{jc}\theta_{rc} \quad \text{for all } i, j$$

where $Q \subseteq R \times C$, which can be written as

$$g(M_{ij}(\theta)) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{C}}\sum_{r\in R^*}\psi_{ir}\theta_{r1} + \frac{1}{\sqrt{R}}\sum_{c\in C^*}\omega_{jc}\theta_{1c} + \sum_{(r,c)\in Q^*}\psi_{ir}\omega_{jc}\theta_{rc} \quad \text{for all } i,j \tag{4}$$

where $R^* \subseteq \{2,\ldots,R\}$, $C^* \subseteq \{2,\ldots,C\}$ and $Q^* \subseteq \{(r,c) : r \in \{2,\ldots,R\},\ c \in \{2,\ldots,C\}\}$. The parameters in each of the four terms in (4) all play a different role.

(1) The first parameter θ_{11} is a constant term.

(2) The θ_{r1} for $r = 2,\ldots,R$ relate exclusively to the rows. Including θ_{r1} in a model means that $\frac{1}{\sqrt{C}}\psi_{ir}\theta_{r1}$ is added to all of the elements in the ith row, so that the modelled probabilities within each row are identical while the rows themselves will differ. For this reason the parameters θ_{r1} for $r = 2,\ldots,R$ will be called the *row-effect parameters.*

(3) The θ_{1c} for $c = 2,\ldots,C$ are to the columns what the θ_{r1} are to the rows and are called *column-effect parameters.*

(4) The inclusion of θ_{rc} in a model leads to $\psi_{ir}\omega_{jc}\theta_{rc}$ being added to the (i,j)th cell for all i,j, i.e. θ_{rc} is added to the cells in the table in the proportions given by the matrix

$$[\psi_{ir}\omega_{jc}]_{i=1,\ldots,R;\ j=1,\ldots,C} = \begin{pmatrix} \psi_{1r}\omega_{1c} & \psi_{1r}\omega_{2c} & \cdots & \psi_{1r}\omega_{Cc} \\ \psi_{2r}\omega_{1c} & \psi_{2r}\omega_{2c} & \cdots & \psi_{2r}\omega_{Cc} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{Rr}\omega_{1c} & \psi_{Rr}\omega_{2c} & \cdots & \psi_{Rr}\omega_{Cc} \end{pmatrix}.$$

This can also be written as either

$$(\omega_{1c}\underline{\psi}_r\ \omega_{2c}\underline{\psi}_r \cdots \omega_{Cc}\underline{\psi}_r) = \underline{\omega}_c' \otimes \underline{\psi}_r$$

or as

$$\begin{pmatrix} \psi_{1r}\underline{\omega}_c' \\ \psi_{2r}\underline{\omega}_c' \\ \vdots \\ \psi_{Rr}\underline{\omega}_c' \end{pmatrix} = \underline{\psi}_r \otimes \underline{\omega}_c'$$

and will be called the $(\underline{\psi}_r, \underline{\omega}_c)$ *interaction matrix.* The corresponding θ_{rc} are called *interaction parameters.*

Higher dimensions. No difficulties arise in extending the above notions to cases in which there are more than two dimensions. We will therefore confine our general discussion to the two dimensional case and consider higher dimensional cases only as they arise in the examples of application. •

3.2 CONSTRUCTING MODEL BASES

The model basis or bases used in the construction of a class of approximating families for a particular table play an important role in determining the fitted probabilities. The choice and construction of the bases is now considered.

At the outset it must be emphasised that there is absolutely no question of using the data to produce an "optimal" basis or bases by, for example, finding eigenvectors, singular-value decompositions, etc, as is done in some exploratory data analysis techniques. The bases should be specified prior to inspection of the data. Selecting bases "suggested" by the data is analogous to postulating a null hypothesis suggested by the same set of data which will be ~~issued~~ used to test this hypothesis. The reason why we must avoid using the data to select a model basis is quite simply that this would lead to fitted models which are "close" to the data but not necessarily close to the operating model. Clearly this would undermine the whole purpose of what one is trying to achieve.

For any table of cross-classified counts, bases are constructed by considering the nature of the variables involved. Generally one constructs a separate model basis for each of the variables. For any given variable, with say L categories, the construction of a model basis is conveniently achieved by choosing $L-1$ orthogonal contrasts between the categories of the variable. (In fact for $L > 2$ one has only to choose $L-2$ contrasts since the remaining contrast can be determined given the others.) Contrasts have been, and still are, used fairly extensively in statistical hypothesis testing and modelling procedures; and the same considerations apply here. However with basis models the contrasts are built directly into the models and one has to *begin* the modelling procedure with all $L-1$ contrasts.

The construction of orthogonal contrasts for a variable will now be illustrated using a number of examples of application. In each case we suggest a model basis for each of the variables, and in some cases two alternative bases are given. In later chapters we will be making use of these data sets (and the bases given below) to illustrate other aspects of model construction and selection.

Before giving details of the individual data sets two general points can be made.

(1) The researcher (the individual(s) who performed or instigated the performance of the experiment and who is ultimately going to use the fitted model) and the statistician should collaborate on the construction of the basis. The researcher often has prior knowledge or expectations about the variables involved which can be usefully incorporated into the model provided these expectations were genuinely held before the data were collected and were not suggested *post hoc*. The statistician's task is to translate this knowledge into contrast vectors which make sense to the researcher. The participation of the researcher in the construction of the basis will lead to the researcher having a greater understanding of the model that is eventually presented to him/her.

(2) In our context there is no equivalent of a "vague prior", i.e. there is no basis which treats all of the cells "equally" and which can be used when there is no prior knowledge.

As regards the actual bases that are proposed it must be stated that these are not the only bases that can or even should be used, and there may well be other more useful bases. In specifying individual bases we will generally not give normalising factors and write for example

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 0 \end{pmatrix}$$

rather than

$$\begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{6} & -1/\sqrt{2} \\ 1/\sqrt{3} & 2/\sqrt{6} & 0 \end{pmatrix}$$

and still refer to the non-normalised basis as a model basis.

THE TREATMENT DATA

Plackett (1974, p.136) gives a data set collected by D.J. Newell. In a clinical trial to compare two analgesics, A and B, 175 patients were randomly allocated to

one of the four sequences AB, BA, AA, BB. The pairs of letters indicate the treatments received by a patient in the first and second periods of the trial, respectively. Each patient was asked to express a preference for the first or second treatment received, with the following results:

		sequence				
		AB	BA	AA	BB	Total
	Prefers first	16	8	10	11	45
preference	Prefers second	4	12	5	6	27
	No preference	20	22	30	31	103
		40	42	45	48	175

Despite the fact that the number of people allocated to each of the sequence categories was determined by a probability process (random allocation), the number of people allocated to each of the sequence categories is not of interest, and we will regard *sequence* as an explanatory variable. *Preference* is regarded as the response to each of the treatment sequences.

We then regard the experiment as being one in which four independent samples were taken; one from each of the sequence categories; where the sample size of each is fixed, having been determined by the allocation process. The operating model is then product-multinomial and the probabilities of interest are those with which an individual will indicate each of the preference categories conditional on the treatment sequence that was administered.

Bases. As the operating model is product-multinomial two bases are required.

We consider choosing a basis for *preference* first. The variable has three categories ("prefers first", "prefers second", "no preference") and so we must choose two contrast vectors. A reasonable contrast is that between the first two categories and the last. The corresponding contrast vector is

$$\begin{pmatrix} 1/2 \\ 1/2 \\ -1 \end{pmatrix}.$$

Having decided on this vector, there is up to scaling factors, only one contrast vector

orthogonal to it, namely

$$\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

In spite of this being the only vector that can be used, it is a particularly convenient one since it contrasts the two categories that were combined in the previous contrast.

The model basis for *preference* we have constructed is thus:

$$\begin{array}{l} \text{prefers first} \\ \text{prefers second} \\ \text{no preference} \end{array} \begin{pmatrix} 1 & 1/2 & 1 \\ 1 & 1/2 & -1 \\ 1 & -1 & 0 \end{pmatrix}.$$

Consider now the choice of a basis for *sequence*. The variable has four categories and so we must choose three contrast vectors. The most noticeable feature of the variable's categories is that the first two, (AB and BA), concern cases where the patients were genuinely given two treatments, which is not the case in the last two categories (AA and BB). This suggests the contrast vector

$$\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}.$$

We now have to choose two more vectors. The natural choice is to use the one vector to contrast AB with BA and the other to contrast AA with BB. The corresponding two vectors are

$$\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

Since the three contrast vectors given above are mutually orthogonal, it follows that a suitable model basis for *sequence* is

$$\begin{array}{l} AB \\ BA \\ AA \\ BB \end{array} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

THE LIZARD DATA

This data is taken from Fienberg (1977) and were originally reported by Schoener (1968). The data were collected by ecologists studying two species of the Anolis Lizards of Bimini. The ecologists were interested in the relationships between the variables that can be used to describe the lizard's habitat – in particular perch height and perch diameter. The counts in the table are of the number of times lizards of each of the two species were observed on each of the perch types.

Anolis Lizards of Bimini

perch height (in feet)	perch diameter (in inches)	species: sagrei	species: distichus
> 4.75	≤ 4.0	32	61
	> 4.0	11	41
≤ 4.75	≤ 4.0	86	73
	> 4.0	35	70
		164	245

Bases. Since each of the variables are binary we use the Hadamard basis H_2 for each. For the two response variables considered jointly the model basis used is $H_4 = H_2 \otimes H_2$, i.e.

height	diameter				
> 4.75	≤ 4.0	1	1	1	1
	> 4.0	1	-1	1	-1
≤ 4.75	≤ 4.0	1	1	-1	-1
	> 4.0	1	-1	-1	1

Of the three contrast vectors in this basis it can be seen that

(1) the first contrasts the two *diameter* categories,

(2) the second contrasts the two *height* categories, while

(3) the third contrasts the *diameter* categories at the same time as it contrasts the

height categories.

The joint basis for the entire eight cell cross-classification is then

$$H_8 = H_2 \otimes H_4 = \frac{1}{\sqrt{2}} \begin{pmatrix} H_4 & H_4 \\ H_4 & -H_4 \end{pmatrix}.$$

Here the first four vectors do not introduce *species* interaction, while the second four do. As regards the first four vectors we can think of them as joining together the corresponding cells in the two *species* columns and then "applying" the model basis H_4 . However for the second four vectors, each of the vectors in H_4 is applied separately to each of the two species which are then contrasted. •

THE CAMP DATA

Bishop *et al* (1975, p.137) introduce a data set first studied by Stouffer *et al* (1949). A sample survey was taken from U.S. army recruits in World War II. The recruits are identified by *race* (black, white), *geographic origin* (North, South) and *location* of current training camp (North, South). The recruits were asked whether they would like to move to another camp, and if so to where they wanted to go. Their answers were categorised as shown.

race		Black				White			
origin		North		South		North		South	
location		North	South	North	South	North	South	North	Sout
preference									
prefer to stay		196	83	261	924	367	346	54	48
prefer to move	North	191	876	122	381	588	874	50	9
	South	36	167	270	788	162	164	176	38
	Undecided	41	153	113	353	191	273	40	9
undecided		52	111	105	272	162	164	40	9
Totals		516	1390	871	2718	1470	1821	360	114

We regard the three variables *race, origin* and *location* as explanatory. The remaining variable, *preference*, is the response. Each member of the population finds

himself in one of the categories of the explanatory variables cross-classification and the only "free" random variable is the respondent's *preference*. The probabilities of interest are of the form

$$\pi_{i(jk\ell)} \quad \text{for} \quad i = 1,\ldots,5;\ j = 1,2;\ k = 1,2;\ \ell = 1,2$$

where this denotes the conditional probability with which an individual, from the jth *race* category, the kth *origin* category and the ℓth *location* category, falls into the ith *preference* category.

Bases. As the operating model is product-multinomial we need separate bases for the explanatory and response variable cross-classifications.

Consider the explanatory variable cross-classification first. In the previous example we had a 2×2 explanatory variable cross-classification for which H_4 was used and which gave us three contrast vectors – namely two main-effect contrast vectors (one for each of the variables) and one interaction contrast vector. We now have a $2 \times 2 \times 2$ explanatory variable cross-classification for which H_8 will be used. In H_8 there are seven contrast vectors, viz.

- 3 main-effect contrast vectors (one each for *race, origin* and *location*)
- $\binom{3}{2} = 3$ two-factor interaction contrast vectors (namely *race* $\star$ *origin*, *race* $\star$ *location* and *origin* $\star$ *location*)
- 1 three-factor interaction contrast vector (namely *race* $\star$ *origin* $\star$ *location*).

The (non-normalised) basis is shown below. In the labelling of the contrast vectors L, O and R stand for *location, origin* and *race* respectively. Notice, for example that the vector labelled L has a +1 in every position corresponding to a *location* value of North and a -1 for all the South values. The $L \star O$ vector, on the other hand has a +1 whenever the *origin* and *location* values are the same and -1 if they are not, and so on.

race	origin	location	·	L	O	L * O	R	L * R	O * R	L * O * R
Black	N	N	1	1	1	1	1	1	1	1
		S	1	-1	1	-1	1	-1	1	-1
	S	N	1	1	-1	-1	1	1	-1	-1
		S	1	-1	-1	1	1	-1	-1	1
White	N	N	1	1	1	1	-1	-1	-1	-1
		S	1	-1	1	-1	-1	1	-1	1
	S	N	1	1	-1	-1	-1	-1	1	1
		S	1	-1	-1	1	-1	1	1	-1

We turn now to the response variable, *preference*, which has an interesting structure among its five categories. Two bases are presented here; these reflect two different ways of looking at the structure.

(1) The first approach is to say that the most notable feature of the categories is that in the first three categories some definite preference is expressed, while in the last two categories some form of undecideness is expressed. Using a contrast between these two groups as a starting point we are led to the basis shown.

prefer to stay		1	1	0	1	0
prefer	North	1	1	0	-1/2	1
to	South	1	1	0	-1/2	-1
move	undecided	1	-3/2	1	0	0
undecided		1	-3/2	-1	0	0

The first contrast vector in this basis contrasts the "decideds" and the "undecideds". The next vector contrasts the two undecided categories. Having thus made allowance for the undecideds, the rest of the contrasts can concentrate on the decideds. The first of these contrasts those who prefer to stay with those who prefer to move; while the second contrasts the prefer-to-move-North with the prefer-to-move-South categories.

(2) A second way to construct a basis for *preference* is to place the emphasis on the order in which the (two) questions were asked and construct the vectors accordingly. One obtains, for example:

prefer to stay		1	-1/4	1	0	0
prefer	North	1	-1/4	-1/3	-1/2	1
to	South	1	-1/4	-1/3	-1/2	-1
move	Undecided	1	-1/4	-1/3	1	0
undecided		1	1	0	0	0

In this case we begin by considering the question "do you want to move?", the answer being one of "no", "yes" or "undecided". Among these three it is reasonable to first contrast the "undecided" category with the first two categories, and then to contrast the "yes" with the "no" category. This is exactly what the first two contrast vectors in the basis are doing. Only once the "do you want to move?" question has been dealt with in this way, is the "where to?" question considered. The categories for answers to the "where to?" question are: "to the North", "to the South" and "undecided". The last two vectors set up the natural contrasts among these categories.

In their analysis of this data set Bishop *et al* (1975) and Goodman (1972) simply discard the two undecided rows. •

THE BEETLE DATA

The data given below were taken from Hewlett and Plackett (1950). They are concerned with the toxicity to the beetle *Tribolium casteneum* of films formed by the insecticide γ -benzene-hexachloride. Six dosage levels of the insecticide were administered, one level to each of a group of either 49 or 50 beetles.

		dose					
		12.08	14.49	16.31	18.13	20.44	22.36
survival	≤ 9 days	20	28	33	30	33	33
	> 9 days	30	21	17	20	17	16
	Total	50	49	50	50	50	49
	Proportion	0.40	0.57	0.66	0.60	0.66	0.67

The dose levels are given by the deposit of 0.1% γ -benzene-hexachloride measured in mg/10 cm^2 . The row marked "proportion" gives the proportion of beetles within each of the groups which died within nine days.

The operating model is clearly product-multinomial, with *dose* the explanatory variable, each of whose categories define a sub-population. Of interest is the probability of beetles dying within the first nine days conditional on each of the dose levels.

Bases. Since the operating model is product-multinomial, we need a separate basis for each of the two variables. *Survival* is binary and so the Hadamard basis H_2 will be used. *Dose*, on the other hand, assumes quantifiable values. This suggests that we use the appropriate orthogonal polynomial basis. Using the original dose levels as given in the table the corresponding orthonormal basis is found to be

$$\begin{array}{c} \begin{array}{cccccc} & x & x^2 & x^3 & x^4 & x^5 \end{array} \\ \begin{pmatrix} .41 & -.67 & .55 & -.27 & .10 & -.02 \\ .41 & -.31 & -.29 & .61 & -.48 & .22 \\ .41 & -.08 & -.46 & .09 & .50 & -.60 \\ .41 & .13 & -.36 & -.43 & .23 & .67 \\ .41 & .37 & .03 & -.42 & -.62 & -.37 \\ .41 & .55 & .52 & .42 & .27 & .10 \end{pmatrix} \end{array} . \qquad (1)$$

Since the *dose* levels are almost equally spaced, a simpler alternative is to use the standard polynomial basis for six equally spaced values. This basis is

$$\begin{pmatrix} 1 & -5 & 5 & -5 & 1 & -1 \\ 1 & -3 & -1 & 7 & -3 & 5 \\ 1 & -1 & -4 & 4 & 2 & -10 \\ 1 & 1 & -4 & -4 & 2 & 10 \\ 1 & 3 & -1 & -7 & -3 & -5 \\ 1 & 5 & 5 & 5 & 1 & 1 \end{pmatrix}. \qquad (2)$$

THE ESKIMO DATA

Bishop *et al* (1975, p.133) analyse a data set reported by Muller and Mayhall (1971) and give the following introduction.

> *"Anthropologists have traditionally used the physical structure of the mouth to study differences among populations and among groups within populations. One often-studied characteristic is the incidence of the morphological trait torus mandibularis, a small protuberance found in the lower jaw at the front of the mouth."*

In the table incidence of *torus mandibularis* is cross-classified by age (six categories) for each of three Eskimo populations.

		population			
incidence	age	Igloolik	Hall Beach	Aleut	Total
present	1-10	5	3	7	15
	11-20	19	8	3	30
	21-30	32	9	6	47
	31-40	31	10	9	50
	41-50	16	8	6	30
	50+	22	6	7	35

absent	1-10	86	39	16	141
	11-20	49	26	20	95
	21-30	38	12	15	65
	31-40	10	4	8	22
	41-50	4	2	7	13
	50+	3	1	4	8

We are provided with the information:

"The first two groups, Igloolik and Hall Beach, are from a pair of villages in the Foxe Basin area of Canada, and the data for these groups were collected by a different investigator than for the third group, the Aleuts from Western Alaska, with a time difference between investigations of about twenty years."

The operating model. We treat *populations* as explanatory, while regarding *incidence* and *age* and response variables. One could regard *age* as explanatory, but we choose not to because the number of people falling into the various age groups is a natural characteristic of each population – and not something which was arbitrarily fixed by the researcher.

Bases. A separate basis is chosen for each of the variables.

Incidence: Since this is a binary variable, the Hadamard basis, H_2 , will be used.

Age: This variable is different from those that have been considered thus far, as its categories are defined by cut-off points on some continuous scale. However because these categories are ordinal and because it is likely that there is some trend in *incidence* with age, we will use an orthogonal polynomial basis. In the absence of any obvious alternative the standard basis derived for equally spaced values is used, namely:

$$\begin{pmatrix} 1 & -5 & 5 & -5 & 1 & -1 \\ 1 & -3 & -1 & 7 & -3 & 5 \\ 1 & -1 & -4 & 4 & 2 & -10 \\ 1 & 1 & -4 & -4 & 2 & 10 \\ 1 & 3 & -1 & -7 & -3 & -5 \\ 1 & 5 & 5 & 5 & 1 & 1 \end{pmatrix}.$$

Population. From what is known about these populations, and the way in which the data were collected, we would want principally to contrast the Igloolik and Hall Beach populations with the Aleut population. The corresponding model basis is

$$\begin{array}{l} \text{Igloolik} \\ \text{Hall Beach} \\ \text{Aleut} \end{array} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix}$$

THE VISION DATA

Two-dimensional tables, with the variable for rows having the same categories as the variable for columns, occur frequently. Such tables may arise in several different ways:

1. in studies where each individual is classified according to the same criterion at two different points in time;

2. when each individual is cross-classified according to two similar categorical variables, such as vision in the left and in the right eye;

3. when each member in a pair of matched individuals, such as father and son, are classified according to some categorical variable, such as political party preferred.

The most prominent feature of this type of cross-classification is that the cells on the main diagonal account for most of the probability. As a representative example of these kind of cross-classifications we look at the data given below, which are based on case-records of the unaided distant vision of male employees aged 30-39 in Royal Ordnance factories in 1943-46 from Stuart (1953). (See Bishop *et al*, (1975, p.284) for the corresponding data set for female employees and for references to previous analyses of the data.) There are two responses, defined by vision in the right and left eyes.

grade of right eye	grade of left eye: highest (1)	second (2)	third (3)	lowest (4)	Totals
highest (1)	821	112	85	35	1053
second (2)	116	494	145	27	782
third (3)	72	151	583	87	893
lowest (4)	43	34	106	331	514
Totals	1052	791	919	480	3242

The operating model is taken to be multinomial since each person tested could have been classified into one of the sixteen cells in the table.

Bases. For this cross-classification, as for most of the type being considered, a single joint basis is constructed for the entire table; not by separately constructing bases for the row and column variables as was the case in the data sets considered hitherto.

Because the cells on the main diagonal account for a large proportion of the probability, in any constrast which involved cells both on and off the diagonal, the off-diagonal cells would be swamped. Thus it is important that the basis be constructed in such a way that the diagonal and off-diagonal cells are contrasted separately.

Consider initially the diagonal cells. Let

$$D = \begin{pmatrix} (1,1) \\ (2,2) \\ (3,3) \\ (4,4) \end{pmatrix}.$$

There are four cells and so three contrast vectors are needed, say $\underline{\omega}_1, \underline{\omega}_2, \underline{\omega}_3 \in \Re^4$. Put

$$\Omega = (\underline{\omega}_1, \underline{\omega}_2, \underline{\omega}_3).$$

We now concentrate on the off-diagonal cells. Consider either of the off-diagonal halves, say the lower.

		left eye			average
		(1)	(2)	(3)	
	(2)	116	-	-	116
right					
	(3)	72	151	-	111.5
eye					
	(4)	43	34	106	61.0

Let

$$L = \begin{pmatrix} (2,1) \\ (3,1) \\ (3,2) \\ (4,1) \\ (4,2) \\ (4,3) \end{pmatrix}$$

be an ordering of these cells. The table has three rows. Let $\underline{\phi}_1, \underline{\phi}_2 \in \Re^6$ be two contrast vectors for contrasting the three row averages. Having contrasted the row averages let $\underline{\phi}_3, \underline{\phi}_4, \underline{\phi}_5 \in \Re^6$ be contrast vectors for contrasting cells within individual rows. The vector $\underline{\phi}_3$ is used to contrast the two cells in the second row, i.e.

$$\underline{\phi}_3 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

while $\underline{\phi}_4$ and $\underline{\phi}_5$ are used to contrast the three cells in the third row. Let

$$\Phi = (\underline{\phi}_1, \ldots, \underline{\phi}_5).$$

Similar considerations hold for the upper off-diagonal cells. It is convenient to re-arrange these cells to a layout which is similar to that of the lower off-diagonal cells, namely,

		right eye			average
		(1)	(2)	(3)	
	(2)	112	-	-	112
left					
	(3)	85	145	-	115
eye					
	(4)	35	27	87	49.7

Define as the upper-half analog to L ,

$$U = \begin{pmatrix} (1,2) \\ (1,3) \\ (2,3) \\ (1,4) \\ (2,4) \\ (3,4) \end{pmatrix}$$

where (i,j) refers to the (i,j)th cell in the *original* cross-classification. Then clearly the contrast vectors used for L , namely Φ , can be used again here for U .

For the complete sixteen-cell table, the cells can be ordered as

$$\begin{pmatrix} U \\ D \\ L \end{pmatrix}$$

and thus a suitable model basis for this ordering is (obtained by normalising)

$$\begin{array}{c} \\ U \\ D \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} \quad & \quad & \quad & \quad & \quad & \quad \end{array} \\ \begin{pmatrix} \underline{1}_6 & 0 & \underline{1}_6 & \Phi & \underline{1}_6 & \Phi \\ \underline{1}_4 & \Omega & -3\underline{1}_4 & 0 & 0 & 0 \\ \underline{1}_6 & 0 & \underline{1}_6 & \Phi & -\underline{1}_6 & -\Phi \end{pmatrix} \end{array}$$

As far as this particular application is concerned it is not clear to us which contrast vectors should be used in Ω and Φ . Although we have previously warned against the practice of looking at the data for guidence in selecting a basis, in the absence of prior expert knowledge, we had no option but to have a "quick look" at the data in this particular case. We observed that the counts in the cells which

involved the lowest grade of vision (for either eye) are generally lower that the others. This suggests that we put

$$\underline{\omega}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -3 \end{pmatrix}$$

and

$$\underline{\phi}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} \quad \text{and hence} \quad \underline{\phi}_2 = \begin{pmatrix} 1 \\ -1/2 \\ -1/2 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The remaining contrasts were chosen essentially arbitrarily, to give

$$\Omega = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \\ -3 & 0 & 0 \end{pmatrix}$$

and

$$\Phi = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & -1/2 & 1 & 0 & 0 \\ 1 & -1/2 & -1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 0 & -2 \\ -1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

3.3 THE MODELLING PROCEDURE

Given a table of cross-classified counts we have seen how to construct suitable model bases and have seen the type of approximating families under consideration. The actual model selection procedure employed will depend on both the link function and the discrepancy function. We begin by describing the common features of the procedures.

Consider a cross-classification involving any number of variables with a multinomial or product-multinomial operating model. Suppose that the cells have been ordered lexicographically and let $[\pi_i]_{i=1,\ldots,L}$ denote the associated operating model probabilities. Let $\mathbf{M}(Q)$ denote an approximating family with members $\underline{M}(\theta) = [M_i(\theta)]_{i=1,\ldots,L}$ where $g(M_i(\theta)) = \Sigma_{q\in Q}\phi_{iq}\theta_q$ and the parameters are subject to the relevant constraints (e.g. $\Sigma_i M_i(\theta) = 1$). The discrepancy between $\underline{M}(\theta)$ and $\underline{\pi}$ is then written $\Delta(\underline{\pi}, \underline{M}(\theta))$ for some discrepancy function $\Delta(\cdot,\cdot)$.

The optimal model within the approximating family $\mathbf{M}(\theta)$ is that which minimises the discrepancy over all members in the family, and the minimum discrepancy parameter, written $\theta^0(Q)$, is defined by

$$\theta^0(Q) = \arg\left\{\min\ \Delta(\underline{\pi}, \underline{M}(\theta))\ :\ \underline{M}(\theta) \in \mathbf{M}(Q)\right\}.$$

This parameter is generally multi-dimensional and its elements will be denoted by $\theta_q^0(Q)$ for $q \in Q$. The reason for writing $\theta_q^0(Q)$ and not just θ_q^0 is that generally, as one might expect, the optimal value of a particular element is in part determined by the other parameters which appear in the model with it, and will not be the same for all approximating families.

The minimum empirical discrepancy parameter estimator within $\mathbf{M}(Q)$ is defined by

$$\hat{\theta}(Q) = \arg\left\{\min\ \Delta(\underline{P}, \underline{M}(\theta))\ :\ \underline{M}(\theta) \in \mathbf{M}(Q)\right\}$$

where $\underline{P}$ is the empirical analog of $\underline{\pi}$, i.e. the vector of sample proportions. The elements of $\hat{\theta}(Q)$ will be denoted by $\hat{\theta}_q(Q)$ for $q \in Q$. The fitted model from

the family $\mathbf{M}(Q)$ is then

$$\underline{M}(\hat{\theta}(Q)) = \sum_{q \in Q} \underline{\phi}_q \hat{\theta}_q.$$

The expected discrepancy of the fitted model $\underline{M}(\hat{\theta}(Q))$ is defined by

$$E_{\underline{\pi}} \Delta\left(\underline{\pi}, \underline{M}(\hat{\theta}(Q))\right) \tag{1}$$

where $E_{\underline{\pi}}$ denotes that the expectation is taken under the operating model. For a given class of approximating families $\{\mathbf{M}(Q) : Q \in \mathbf{S}(L)\}$ the optimal fitted model is that which minimises (an estimator of) (1) over all $Q \in \mathbf{S}(L)$.

As was stated in Chapter 1 the model selection problem could involve considerable computation, but for some (link function, discrepancy function) pairs it is possible to exploit the orthogonality of the contrast vectors to achieve either or both of the following properties:

(i) the estimate of a parameter is the same no matter what other parameters are in the model, and

(ii) each parameter contributes separately to the estimated expected discrepancy.

In this thesis we will consider two discrepancy functions which we will now introduce.

Discrepancy functions. The first of the two discrepancy functions is based on "squared errors" and is called the *Gaussian discrepancy function* by Linhart and Zucchini (1986a). For a multinomial operating model $[\pi_i]_{i=1,\ldots,L}$ and model $[M_i(\theta)]_{i=1,\ldots,L}$ the Gaussian discrepancy is defined by

$$\sum_{i=1}^{L} (\pi_i - M_i(\theta))^2$$

while for a product-multinomial operating model $[\pi_{i(j)}]_{i=1,\ldots,R;\, j=1,\ldots,C}$ this discrepancy is defined by

$$\sum_{j=1}^{C} \sum_{i=1}^{R} (\pi_{i(j)} - M_{i(j)}(\theta))^2.$$

The second discrepancy function is the well-known ***Kullback-Leibler discrepancy function***, which has the property that minimum empirical Kullback-Leibler discrepancy parameter estimators are, in general, equal to maximum likelihood estimators.

For the multinomial case this discrepancy is defined by

$$\sum_{i=1}^{L}(E_{\underline{\pi}}n_i)\log\left(\pi_i/M_i(\theta)\right) = n_+\sum_{i=1}^{L}\pi_i\log(\pi_i/M_i(\theta))$$

while for the product-multinomial case this discrepancy is defined by

$$\sum_{j=1}^{C}\sum_{i=1}^{R}(E_{\underline{\pi}}n_{ij})\log\left(\pi_{i(j)}/M_{i(j)}(\theta)\right) = \sum_{j=1}^{C}n_{+j}\sum_{i=1}^{R}\pi_{i(j)}\log(\pi_{i(j)}/M_{i(j)}(\theta)).$$

Both the Gaussian and the Kullback-Leibler discrepancy functions contain terms which remain constant for all approximating families and are thus inessential for the purposes of comparing competing fitted models (for a given data set). The essential parts of the four discrepancy functions defined above are, respectively

- $-2\Sigma_i\pi_i M_i(\theta) + \Sigma_i(M_i(\theta))^2$
- $-2\Sigma_j\Sigma_i\pi_{i(j)}M_{i(j)}(\theta) + \Sigma_j\Sigma_i(M_{i(j)}(\theta))^2$
- $-n_+\Sigma_i\pi_i\log\ M_i(\theta)$
- $-\Sigma_j n_{+j}\Sigma_i\pi_{i(j)}\log\ M_{i(j)}(\theta)$

Frequently there is no need to differentiate between the actual- and the essential discrepancy function and then we will simply use the term discrepancy function in both cases.

The two (link function, discrepancy function) pairs that we will consider are

(1) the identity link with the Gaussian discrepancy function, and

(2) the log (natural logarithm) link with the Kullback-Leibler discrepancy function

the first in Chapter 4 (and again in a slightly different context in Chapter 6), and the second in Chapter 5. For the purposes of comparison, the data sets and the model bases that were introduced in Section 3.2 are used as illustrative examples in Chapter 4 and again in Chapter 5. •

CHAPTER 4
LINEAR MODELS

In this chapter we concentrate on linear models and the Gaussian discrepancy. In the first section a model selection algorithm is developed; in the second section linear models are fitted to each of the data sets introduced in Section 3.2.

4.1 THE MODEL SELECTION PROCEDURE

As pointed out in the previous section, model selection requires considerable computation unless we are able to exploit the orthogonality of the contrast vectors. For the case under consideration, that of the identity link and the Gaussian discrepancy, the orthogonality can be fully exploited and it is possible to construct a very simple algorithm for selecting models.

The multinomial and product-multinomial cases will be considered separately. The results derived for the two cases are similar.

A. MULTINOMIAL OPERATING MODELS

Consider a cross-classification, with any number of variables, whose operating model is multinomial. Suppose that the cells are arranged into a single vector of length L and let $[\pi_i]_{i=1,\ldots,L}$ be the operating model probabilities. In addition let $\Phi = [\phi_{iq}]_{i=1,\ldots,L;\ q=1,\ldots,L}$ be a model basis for the table. Then the class of approximating families under consideration has members

$$\mathbf{M}(Q) = \left\{ [M_i(\theta)]_{i=1,\ldots,L} \ : \ M_i(\theta) = \sum_{q=1}^{L} \phi_{iq}\theta_q;\ \sum_i M_i(\theta) = 1 \right\}.$$

The Gaussian discrepancy function for $\underline{M}(\theta) = [M_i(\theta)]_i \in \mathbf{M}(Q)$ is given by

$$\Delta(\underline{\pi}, \underline{M}(\theta)) = \sum_{i=1}^{L} (\pi_i - M_i(\theta))^2. \tag{1}$$

Since the vectors $\underline{\phi}_1, \ldots, \underline{\phi}_L$ are orthonormal the right hand side of (1) can be written as

$$\sum_{i=1}^{L} \pi_i^2 - 2 \sum_{q \in Q} \theta_q \left(\sum_{i=1}^{L} \phi_{iq}\pi_i \right) + \sum_{q \in Q} \theta_q^2 \tag{2}$$

which is the form most often used in deriving orthogonality properties.

The first of the orthogonality properties concerns the minimum discrepancy parameters. The (multi-dimensional) minimum discrepancy parameter $\theta^0(Q)$ is found by minimising (2), with respect to the θ_q for $q \in Q$, subject to the constraint $\Sigma_i M_i(\theta) = 1$. The method of Lagrangian multipliers is used and it is found that a solution can only be obtained provided $1 \in Q$, i.e. provided θ_1 is in the model. (This condition will be looked at later.)

Define

$$G(\theta, \lambda) = \Delta(\underline{\pi}, \underline{M}(\theta)) + \lambda(\Sigma_i M_i(\theta) - 1)$$

where λ is a Lagrangrian multiplier. The $\theta_q^0(Q)$ are found by simultaneously solving the equations

$$\left.\begin{aligned} \frac{\partial G(\theta,\lambda)}{\partial \theta_q} &= 0 \quad \text{for } q \in Q \\ \frac{\partial G(\theta,\lambda)}{\partial \lambda} &= 0. \end{aligned}\right\}$$

Now

$$\begin{aligned} \sum_i M_i(\theta) &= \sum_i \left(\sum_{q \in Q} \phi_{iq} \theta_q \right) \\ &= \sum_{q=1}^{L} \theta_q \left(\sum_i \phi_{iq} \right) = \sqrt{L}\theta_1 \end{aligned}$$

so that

$$G(\theta, \lambda) = \sum_{i=1}^{L} \pi_i^2 - 2 \sum_{q \in Q} \theta_q \left(\sum_{i=1}^{L} \phi_{iq} \pi_i \right) + \sum_{q \in Q} \theta_q^2 + \lambda(\sqrt{L}\theta_1 - 1).$$

Hence

$$\frac{\partial G(\theta,\lambda)}{\partial \theta_q} = \begin{cases} -2\frac{1}{\sqrt{L}} + 2\theta_1 + \lambda\sqrt{L} & \text{if } q = 1 \\ -2\Sigma_i \phi_{iq} \pi_i + 2\theta_q & \text{otherwise} \end{cases}$$

while

$$\frac{\partial G(\theta,\lambda)}{\partial \lambda} = \sqrt{L}\theta_1 - 1.$$

Provided $1 \in q$, $\frac{\partial G(\theta,\lambda)}{\partial \theta_1} = 0$ and $\frac{\partial G(\theta,\lambda)}{\partial \lambda} = 0$ can be solved simultaneously to give $\lambda = 0$. The equation $\frac{\partial G(\theta,\lambda)}{\partial \lambda} = 0$ is then redundant and the resulting system can be written as

$$-2\Sigma_i \phi_{iq} \pi_i + 2\theta_q = 0 \quad \text{for all } q \in Q$$

which has the unique solution

$$\theta_q^0(Q) = \Sigma_i \phi_{iq} \pi_i \quad \text{for } q \in Q. \tag{3}$$

We note immediately that the optimal parameters are independent of Q, i.e. the optimal value of any parameter θ_q is independent of what other parameters are included in the model with it. This property is extremely useful. Among other things it means that instead of having to find the optimal values anew for each different approximating family, one has only to find the optimal values of $\theta_1, \ldots, \theta_L$ once. (This can be done by finding the minimum discrepancy parameters for the saturated approximating family.) If $\theta_1^0, \ldots, \theta_L^0$ denote these optimal values then the optimal model from the approximating family $\mathbf{M}(Q)$ (for any Q) is given directly by

$$M_i(\theta^0) = \sum_{q \in Q} \phi_{iq} \theta_q^0 \quad \text{for } i = 1, \ldots, L.$$

The minimum empirical discrepancy estimator $\hat{\theta}(Q)$ is the empirical analog of $\theta^0(Q)$. The estimates share the orthogonality property so that the parameters $\theta_1, \ldots, \theta_L$ have to be estimated only once. This of course considerably reduces the computational load. The elements of $\hat{\theta}(\{1, \ldots, L\})$ are

$$\begin{aligned} \hat{\theta}_q &= \Sigma_i \phi_{iq} P_i \\ &= (\underline{\phi}_q \cdot \underline{P}) \quad \text{for } q = 1, \ldots, L. \end{aligned}$$

Each $\hat{\theta}_q$ is also the unique minimum variance unbiased (UMVU) estimate and the maximum likelihood estimate of θ_q^0.

The (only) fitted model from the approximating family $\mathbf{M}(Q)$ that will be considered is

$$M_i(\hat{\theta}) = \sum_{q \in Q} \phi_{iq} \hat{\theta}_q \quad \text{for } i = 1, \ldots, L.$$

We now consider the inclusion of θ_1 in all models. Since $\theta_1^0 = \frac{1}{\sqrt{L}}$ is constant, $\hat{\theta}_1 = \frac{1}{\sqrt{L}}$ and all fitted models which contain $\hat{\theta}_1$ can be written as

$$M_i(\hat{\theta}) = \frac{1}{L} + \sum_{q \in Q^*} \phi_{iq} \hat{\theta}_q \quad \text{for } i = 1, \ldots, L,$$

where the first term assigns equal probability to each of the cells. As has previously been stated, this first term can be thought of as providing a basic model which can be made more complex by the inclusion of any of the other parameters $(\theta_2, \ldots, \theta_L)$ into the model. The inclusion of θ_1

(1) does not introduce any variance into the model (since $\hat{\theta}_1$ is constant); and

(2) guarantees that the fitted probabilities sum to ~~zero~~ one.

The second of the orthogonality properties involves the expected discrepancy. The expected discrepancy of the fitted model $\underline{M}(\hat{\theta})$ from the family $\mathbf{M}(Q)$ is

$$\begin{aligned} E_{\underline{\pi}}\Delta(\underline{\pi}, \underline{M}(\hat{\theta})) &= E_{\underline{\pi}}\left\{\sum_{i=1}^{L}\pi_i^2 - 2\sum_{q\in Q}\hat{\theta}_q\theta_q^0 + \sum_{q\in Q}\hat{\theta}_q^2\right\} \\ &= \sum_{i=1}^{L}\pi_i^2 + \sum_{q\in Q}\{E_{\underline{\pi}}(\hat{\theta}_q^2) - 2(E_{\underline{\pi}}\hat{\theta}_q)^2\} \\ &= \sum_{i=1}^{L}\pi_i^2 + \sum_{q\in Q}\{2\mathrm{var}_{\underline{\pi}}\hat{\theta}_q - E_{\underline{\pi}}(\hat{\theta}_q^2)\}. \end{aligned}$$

The first term is inessential. The second can be written as

$$\sum_{q\in Q} f(\hat{\theta}_q)$$

where

$$f(\hat{\theta}_q) = 2\,\mathrm{var}_{\underline{\pi}}\hat{\theta}_q - E_{\underline{\pi}}(\hat{\theta}_q^2).$$

It can be seen that each parameter $\hat{\theta}_q$ in the model makes a separate contribution to the expected discrepancy. This greatly simplifies the model selection procedure.

If $f(\hat{\theta}_q)$ is negative, the inclusion of $\hat{\theta}_q$ in the model actually leads to a *decrease* in the expected discrepancy, and so it should be included in the fitted model. Conversely, $f(\hat{\theta}_q)$ being positive implies that $\hat{\theta}_q$ should not be included. The optimal fitted model is thus

$$M_i(\hat{\theta}) = \frac{1}{L} + \sum_{q\in Q^0}\phi_{iq}\hat{\theta}_q \quad \text{for } i = 1, \ldots, L$$

where

$$Q^0 = \{q \in \{2, \ldots, L\} \ : \ f(\hat{\theta}_q) < 0\}.$$

In practice the $f(\hat{\theta}_q)$ cannot be evaluated and they have to be estimated, by say $c(\hat{\theta}_q)$. We will use the UMVU estimator

$$c(\hat{\theta}_q) = 2\,\hat{\text{var}}\,\hat{\theta}_q - \hat{\theta}_q^2$$

where

$$\begin{aligned}\hat{\text{var}}\,\hat{\theta}_q &= \sum_{i=1}^{L} \phi_{iq}^2 \left(\frac{P_i(1-P_i)}{n_+ - 1}\right) + \sum_i \sum_{j \neq i} \phi_{iq}\phi_{jq}\left(-\frac{P_i P_j}{n_+ - 1}\right) \\ &= \frac{1}{n_+ - 1}\left(\sum_i \phi_{iq}^2 P_i - (\sum_i \phi_{iq} P_i)^2\right).\end{aligned}$$

The computations required for model selection are summarised in the following algorithm:

(1) For $q = 2, \ldots, L$ compute

$$\begin{aligned}\hat{\theta}_q &= \Sigma_i \phi_{iq} P_i \\ \hat{\text{var}}\,\hat{\theta}_q &= \frac{1}{n_+ - 1}(\Sigma_i \phi_{iq}^2 P_i - \hat{\theta}_q^2) \\ c(\hat{\theta}_q) &= 2\,\hat{\text{var}}\,\hat{\theta}_q - \hat{\theta}_q^2.\end{aligned}$$

(2) Put

$$Q^0 = \{q \in \{2, \ldots, L\} \ : \ c(\hat{\theta}_q) < 0\}.$$

(3) The fitted model which is estimated to be optimal is

$$M_i(\hat{\theta}) = \frac{1}{L} + \sum_{q \in Q^0} \phi_{iq}\hat{\theta}_q \quad \text{for } i = 1, \ldots, L.$$

Note that $c(\hat{\theta}_q) < 0$ iff

$$|\hat{\theta}_q| \Big/ \sqrt{\hat{\text{var}}\,\hat{\theta}_q} > \sqrt{2}. \tag{4}$$

and thus the selection procedure reduces to the simple rule:

$\hat{\theta}_q$ is to be included in the final fitted model iff (4) holds.

The idea that a parameter is considered worthwhile incorporating into a fitted model only if its absolute value is large relative to its standard deviation has an immediate intuitive appeal.

Further insight can be gained from considerations of the discrepancy components. It is not difficult to show that

(1) the expected discrepancy due to approximation of a family $\mathbf{M}(Q)$ is

$$\Delta(\underline{\pi}, \underline{M}(\theta^0)) = \sum_{q=1}^{L} (\theta_q^0)^2 - \sum_{q \in Q} (\theta_q^0)^2,$$

(2) the expected discrepancy due to estimation of the fitted model $\underline{M}(\hat{\theta}) \in \mathbf{M}(Q)$ is

$$E_{\underline{\pi}} \Delta(\underline{M}(\theta^0), \underline{M}(\hat{\theta})) = \sum_{q \in Q} \text{var}_{\underline{\pi}} \hat{\theta}_q$$

and that the total expected discrepancy between $\underline{M}(\hat{\theta})$ and $\underline{\pi}$ is the sum of the above two components, i.e.

$$E_{\underline{\pi}} \Delta(\underline{\pi}, \underline{M}(\hat{\theta})) = \Delta(\underline{\pi}, \underline{M}(\theta^0)) + E_{\underline{\pi}} \Delta(\underline{M}(\theta^0), \underline{M}(\hat{\theta})).$$

From this one sees immediately that the larger $|\theta_q^0|$ is, the greater will be the decrease in the expected discrepancy, while the larger $\text{var}_{\underline{\pi}} \hat{\theta}_q$ is, the greater the increase in the expected discrepancy.

As a final point note that the critical value that the Gaussian discrepancy gives us to compare $|\hat{\theta}_q| / \sqrt{\hat{\text{var}}\, \hat{\theta}_q}$ with, is $\sqrt{2}$. Increasing this critical value makes it more difficult to include more estimated parameters into the fitted model and amounts to giving more weight to the expected discrepancy due to estimation than to the discrepancy due to approximation. Decreasing the critical value has the reverse effect. One can of course alter the critical value; in effect this changes the discrepancy function.

B. PRODUCT-MULTINOMIAL OPERATING MODELS

We now turn attention to product-multinomial operating models. The models, the selection procedure and the interpretations are all quite similar to those given for multinomial operating models and so only the essentials will be listed here.

Consider a general product-multinomial operating model, and let $[\pi_{i(j)}]_{i=1,\ldots,R;\, j=1,\ldots,C}$ denote the indexed probabilities, where both the row and column variables may in fact be multi-dimensional. A typical approximating family is then

$$\mathbf{M}(Q) = \left\{ [M_{i(j)}(\theta)]_{i=1,\ldots,R;\, j=1,\ldots C} \;:\; M_{i(j)}(\theta) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc}\,; \right.$$
$$\left. \sum_i M_{i(j)}(\theta) = 1 \quad \text{for } j = 1,\ldots,C \right\}$$

for some $Q \subseteq R \times C$. The discrepancy between $M(\theta) = [M_{i(j)}(\theta)]_{i,j}$ and $\pi = [\pi_{i(j)}]_{i,j}$ is defined by

$$\Delta(\pi, \underline{M}(\theta)) = \sum_{j=1}^{C}\sum_{i=1}^{R}(\pi_{i(j)} - M_{i(j)}(\theta))^2.$$

The minimum discrepancy parameters are again found using the method of Lagrangian multipliers. The objective function now contains C multipliers, because of the C restrictions $\Sigma_i M_{i(j)}(\theta) = 1$, and in solving the minimisation problem it is convenient to assume that $Q \supseteq \{(1,c) \;:\; c = 1,\ldots,C\}$. This leads to

$$\theta^0_{rc} = \Sigma_j \omega_{jc}(\Sigma_i \psi_{ir}\pi_{i(j)}) \quad \text{for all } (r,c) \in Q, \tag{5}$$

and in particular

$$\theta^0_{1c} = \begin{cases} \sqrt{C/R} & \text{if } c = 1 \\ 0 & \text{otherwise.} \end{cases}$$

If

$$R^* \times C = \{(r,c) \;:\; r \in \{2,\ldots,R\},\; c \in \{1,\ldots,C\}\}$$

then the class of approximating families considered is

$$\{M(Q) \;:\; Q \in \{(1,1) \cup (R^* \times C)\}\} \tag{6}$$

and the optimal model in $M(Q)$ can be written as

$$M_{i(j)}(\theta^0) = \frac{1}{R} + \sum_{(r,c)\in Q^*} \psi_{ir}\omega_{jc}\theta^0_{rc} \quad \text{for all } i,j$$

where $Q^* \subseteq R^* \times C$ and the form of the θ^0_{rc} is as given in (5). Note that the first term in the model given above now assigns equal probability to each of the R cells in each column.

The parameters (5) are estimated by

$$\hat{\theta}_{rc} = \Sigma_j \omega_{jc}(\Sigma_i \psi_{ir} P_{i(j)}) \quad \text{for all } (r,c) \in Q$$

to give the fitted model $[M_{i(j)}(\hat{\theta})]_{i,j} \in \mathbf{M}(Q)$ where

$$M_{i(j)}(\theta) = \frac{1}{R} + \sum_{(r,c)\in Q^*} \psi_{ir}\omega_{jc}\theta_{rc} \quad \text{for all } i,j$$

whose expected discrepancy is given by

$$E_\pi \Delta(\pi, M(\hat{\theta})) = \sum_j \sum_i \pi^2_{i(j)} + \sum_{(r,c)\in Q} \{2 \operatorname{var}_\pi \hat{\theta}_{rc} - E_\pi(\hat{\theta}^2_{rc})\}$$

where

$$\begin{aligned} \operatorname{var}_\pi \hat{\theta}_{rc} &= \operatorname{var}_\pi(\Sigma_j \omega_{jc}(\Sigma_i \psi_{ir} P_{i(j)})) \\ &= \Sigma_j \omega^2_{jc} \operatorname{var}_{\underline{\pi}(j)}(\Sigma_i \psi_{ir} P_{i(j)}). \end{aligned}$$

The UMVU estimator of $\{2 \operatorname{var}_\pi \hat{\theta}_{rc} - E_\pi(\hat{\theta}^2_{rc})\}$ is

$$c(\hat{\theta}_{rc}) = 2 \hat{\operatorname{var}}\, \hat{\theta}_{rc} - \hat{\theta}^2_{rc}$$

where

$$\hat{\operatorname{var}}\, \hat{\theta}_{rc} = \Sigma_j \omega^2_{jc} \frac{1}{n_{+j} - 1} \{\Sigma_i \psi^2_{ir} P_{i(j)} - (\Sigma_i \psi_{ir} P_{i(j)})^2\}.$$

The steps used to select the fitted model from the class of approximating families (6) are summarised in the following algorithm:

(1) For $r = 2, \ldots, R$ and $c = 1, \ldots, C$ compute

$$\hat{\theta}_{rc} = \Sigma_j \omega_{jc} (\Sigma_i \psi_{ir} P_{i(j)})$$

$$\hat{\text{var}}\, \hat{\theta}_{rc} = \Sigma_j \omega_{jc}^2 \frac{1}{(n_{+j} - 1)} \{\Sigma_i \psi_{ir}^2 P_{i(j)} - (\Sigma_i \psi_{ir} P_{i(j)})^2\}$$

$$c(\hat{\theta}_{rc}) = 2\, \hat{\text{var}}\, \hat{\theta}_{rc} - \hat{\theta}_{rc}^2$$

(2) Put

$$Q^0 = \{(r, c) \in R^* \times C \ : \ c(\hat{\theta}_{rc}) < 0\}.$$

(3) The fitted model which is estimated to be optimal is given by

$$M_{i(j)}(\hat{\theta}) = \frac{1}{R} + \sum_{(r,c) \in Q^0} \psi_{ir} \omega_{jc} \hat{\theta}_{rc} \quad \text{for all } i, j. \bullet$$

4.2 EXAMPLES

Linear models are now fitted to each of the data sets that were introduced in Section 3.2, using the model bases that were given there.

THE TREATMENT DATA

Consider the treatment data introduced in Section 3.2. The relevant sample proportions, expressed as percentages, are shown:

	sequence			
preference	AB	BA	AA	BB
prefers first	40.00	19.05	22.22	22.92
prefers second	10.00	28.57	11.11	12.50
no preference	50.00	52.38	66.67	64.58
	100.00	100.00	100.00	100.00

The model basis to be used for the variable, *preference*, is

$$\begin{array}{l} \text{prefers first} \\ \text{prefers second} \\ \text{no preference} \end{array} \begin{pmatrix} 1/\sqrt{3} & 0.5/\sqrt{1.5} & 1/\sqrt{2} \\ 1/\sqrt{3} & 0.5/\sqrt{1.5} & -1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{1.5} & 0 \end{pmatrix}$$

while for *sequence* it is

$$\begin{array}{l} \text{AB} \\ \text{BA} \\ \text{AA} \\ \text{BB} \end{array} \begin{pmatrix} 1/2 & 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 1/2 & -1/\sqrt{2} & 0 \\ 1/2 & -1/2 & 0 & 1/\sqrt{2} \\ 1/2 & -1/2 & 0 & -1/\sqrt{2} \end{pmatrix}.$$

Let $\pi_{i(j)}$ denote the probability with which a patient given the jth treatment sequence $j = 1,\ldots,4$ indicates the ith preference category. The fitted models considered are of the form

$$M_{i(j)}(\hat{\theta}) = \frac{1}{3} + \sum_{r\in A}\sum_{c\in B} \psi_{ir}\omega_{jc}\hat{\theta}_{rc} \quad \text{for all } i,j$$

where $[\psi_{ir}]_{i,r}$ and $[\omega_{jc}]_{j,c}$ are the model bases given above, and

$$\hat{\theta}_{rc} = \sum_{j=1}^{4} \omega_{jc} \left(\sum_{i=1}^{3} \psi_{ir} P_{i(j)} \right) \quad \text{for all } r \in A;\ c \in B$$

where A and B are two sets, $A \subseteq \{2,3\}$, $B \subseteq \{1,\ldots,4\}$.

We consider first the $\hat{\theta}_{2c}$ for $c = 1,\ldots,4$ and then the $\hat{\theta}_{3c}$ for $c = 1,\ldots,4$.

The $\hat{\theta}_{2c}$ all involve the row contrast vector

$$\underline{\psi}_2 = \frac{1}{\sqrt{1.5}} \begin{pmatrix} 0.5 \\ 0.5 \\ -1 \end{pmatrix}$$

which contrasts the first two preference categories, "prefers first" and "prefers second", with the "no preference" category. The parameters are

$$\hat{\theta}_{2c} = \frac{1}{\sqrt{1.5}} \sum_{j=1}^{4} \omega_{jc} \left(\frac{P_{1(j)} + P_{2(j)}}{2} - P_{3(j)} \right) \quad \text{for } c = 1,\ldots,4.$$

If we define

$$\mathrm{PON}(j) = \frac{P_{1(j)} + P_{2(j)}}{2} - P_{3(j)} \quad \text{for } j = 1,\ldots,4$$

as the "preference-or-not" contrast for the jth treatment combination then we can write

$$\hat{\theta}_{21} = \frac{1}{\sqrt{6}} \mathrm{PON}(+)$$

$$\hat{\theta}_{22} = \frac{1}{\sqrt{6}} ((\mathrm{PON}(1) + \mathrm{PON}(2)) - (\mathrm{PON}(3) + \mathrm{PON}(4))).$$

The interaction matrix of $\hat{\theta}_{2c}$ (which determines the proportions according to which $\hat{\theta}_{2c}$ is added to each of the cells in forming the modelled probabilities) is the 3×4 matrix

$$\underline{\psi}_2 \times \underline{\omega}'_c = \frac{1}{\sqrt{1.5}} \begin{pmatrix} 0.5 & \underline{\omega}'_c \\ 0.5 & \underline{\omega}'_c \\ -1 & \underline{\omega}'_c \end{pmatrix}.$$

In each of these matrices, for $c = 1,\ldots,4$, the first two rows are identical so that including any of these parameters will mean that the modelled probabilities in the first two rows will be the same, but different from the corresponding modelled probability in the third row.

The parameter estimates, their estimated standard deviations and their contributions to the estimated expected discrepancy are shown below. All entries in the table have been multiplied by 10^3 .

	c = 1	c = 2	c = 3	c = 4
$\hat{\theta}_{2c} \times 10^3$	-614.2	176.8	20.6	-18.0
$\sqrt{\text{var}\,\hat{\theta}_{2c}} \times 10^3$	91.7	91.7	96.8	86.2
$c(\hat{\theta}_{2c}) \times 10^3$	-360.4	-14.4	18.3	14.6

Since $c(\hat{\theta}_{21})$ and $c(\hat{\theta}_{22})$ are negative the two corresponding parameters should be included in the final fitted model. Including $\hat{\theta}_{21}$ in the model means adding

$$\hat{\theta}_{21}(\underline{\psi}_2 \otimes \underline{\omega}_1') = \frac{-0.6142}{\sqrt{6}} \begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ -1 & -1 & -1 & -1 \end{pmatrix}$$

to the modelled cell probabilities. This simply models a difference between the average of the first two row marginals and the last row marginal. The fact that $\hat{\theta}_{21}$ is negative means, in view of the nature of its interaction matrix, that more patients indicated "no preference" rather than specifying a definite preference. This is not surprising since two of the four treatment sequences involve giving the same treatment twice.

As for $\hat{\theta}_{22}$, its inclusion means adding

$$\frac{.1768}{\sqrt{6}} \begin{pmatrix} 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ -1 & -1 & 1 & 1 \end{pmatrix}$$

to the table of modelled cell probabilities. This interaction matrix takes the contrast between the first two and the last preference categories and contrasts this between the two sets of treatment sequences { AB, BA } and { AA, BB } . That $\hat{\theta}_{22}$ is positive indicates that the probabilities in the two groups of cells corresponding to patients who either

- received two different treatments and expressed a definite preference, or
- received the same treatment twice and expressed "no preference"

will be increased, while the remaining modelled probabilities will be decreased. The inclusion of $\hat{\theta}_{22}$ suggests that patients who received the same treatment twice were more likely to indicate "no preference" than those who received two different treatments. Again this is what one would expect.

The remaining two parameters, which respectively involve contrasting the "preference-or-not" contrast with

- a contrast between AB and BA
- a contrast between AA and BB

are not considered worthwhile including. This indicates that whether patients had a definite preference or not is unaffected by whether they were given AB rather than BA or given AA rather than BB. Once again this is what we would expect.

We now consider the $\hat{\theta}_{3c}$ for $c = 1, \ldots, 4$ whose interaction matrices are

$$\underline{\psi}_3 \otimes \underline{\omega}_c' = \frac{1}{\sqrt{2}} \begin{pmatrix} \underline{\omega}_c' \\ -\underline{\omega}_c' \\ \underline{0}' \end{pmatrix} \qquad \text{for } c = 1, \ldots, 4$$

so that the parameters can be used to introduce a difference in modelled cell probabilities between corresponding cells in the first and second rows. The $\hat{\theta}_{23}$ are defined by

$$\hat{\theta}_{3c} = \frac{1}{\sqrt{2}} \sum_{j=1}^{4} \omega_{jc} (P_{1(j)} - P_{2(j)}) \quad \text{for } c = 1, \ldots, 4$$

where $(P_{1(j)} - P_{2(j)})$ is the difference between the relative sampled proportions in the "prefer first" and "prefer second" cells for the jth treatment sequence.

The analog of the table given above for the $\hat{\theta}_{2c}$ is shown.

	c = 1	c = 2	c = 3	c = 4
$\hat{\theta}_{3c} \times 10^3$	148.5	-3.7	197.6	3.5
$\sqrt{\text{var}\,\hat{\theta}_{3c}} \times 10^3$	67.7	67.6	74.0	60.4
$c(\hat{\theta}_{3c}) \times 10^3$	-12.9	9.1	-28.1	7.3

The selection criterion indicates that $\hat{\theta}_{31}$ and $\hat{\theta}_{33}$ should be included in the fitted model. The inclusion of $\hat{\theta}_{31}$ means adding 0.1485 to each of the cells in the first row and subtracting the same quantity from the cells in the second row. That $\hat{\theta}_{31}$ is included (and is positive) suggests that patients have a tendency to prefer the first treatment they were given. On the other hand the exclusion of $\hat{\theta}_{32}$ suggests that there is no difference in this trend between the first two and the last two treatment sequences. Finally, the inclusion of $\hat{\theta}_{33} = 0.1976$ with its interaction matrix

$$\frac{1}{2}\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

suggests that

- of those given AB most prefer the first, viz A
- of those given BA most prefer the second, viz A,

which can be interpreted as indicating that of those who received two different treatments there is a tendency for patients to prefer treatment A to B irrespective of the order in which the two treatments are given.

The final fitted model then contains only the four parameters, $\hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{31}$ and $\hat{\theta}_{33}$. The fitted conditional probabilities are shown.

		sequence			
		AB	BA	AA	BB
preference	Prefers first	39.5	19.8	22.4	22.4
	Prefers second	9.3	29.0	11.9	11.9
	No preference	51.2	51.2	65.6	65.6
		100.0	100.0	100.0	100.0

The remaining data sets will be discussed in less detail. It will be convenient to represent the quantities of interest in the model selection process in two tables. The first, which will be called simply *the parameter table*, will give the parameter estimates with their estimate standard deviations shown in brackets. The second, called *the criterion table*, will give the contributions to the estimated expected discrepancy. (We will adopt the convention throughout the thesis that all the entries in these tables will have been multiplied by 10^3 .)

The parameter and criterion tables for the treatment data are:

The parameter table

		sequence			
		$\underline{\omega}_1$	$\underline{\omega}_2$	$\underline{\omega}_3$	$\underline{\omega}_4$
preference	$\underline{\psi}_2$	-614.20	176.79	20.62	-18.04
		(91.67)	(91.67)	(96.80)	(86.24)
	$\underline{\psi}_3$	148.51	-3.72	197.62	3.47
		(67.55)	(67.55)	(74.00)	(60.41)

The criterion table

		sequence			
		$\underline{\omega}_1$	$\underline{\omega}_2$	$\underline{\omega}_3$	$\underline{\omega}_4$
preference	$\underline{\psi}_2$	-360.4	-14.4	18.3	14.6
	$\underline{\psi}_3$	-12.9	9.1	-28.1	7.3

THE LIZARD DATA

Consider the lizard data introduced in Section 3.2 for which the sampled proportions, expressed as percentages, are shown.

perch height (in feet)	perch diameter (in inches)	species: sagrei	species: distichus
> 4.75	≤ 4.0	19.5	24.9
	> 4.0	6.7	16.7
≤ 4.75	≤ 4.0	52.4	29.8
	> 4.0	21.3	28.6
		100	100

Although there are two response variables, it is convenient to regard the *diameter* × *height* cross-classification as a single classification. We will let $\pi_{i(j)}$ denote the probability with which a lizard of the jth *species* category $(j = 1, 2)$ is found on a perch which falls into the ith cell of the *diameter* × *height* classification $(i = 1, \ldots, 4)$.

For the *diameter* × *height* classification the model basis used is H_4 :

height	diameter				
> 4.75	≤ 4.0	1	1	1	1
	> 4.0	1	-1	1	-1
≤ 4.75	≤ 4.0	1	1	-1	-1
	> 4.0	1	-1	-1	1

If $\underline{\psi}_r$ denotes the rth column of this basis, then $\underline{\psi}_2, \underline{\psi}_3$ and $\underline{\psi}_4$ are respectively *diameter, height* and *diameter* $\star$ *height* contrast vectors.

For the explanatory variable, *species*, H_2 is used.

The saturated fitted model is then

$$M_{i(j)}(\theta) = \frac{1}{4} + \sum_{r=2}^{4} \sum_{c=1}^{2} \psi_{ir} \omega_{jc} \hat{\theta}_{rc} \quad \text{for } i = 1, \ldots, 4;\ j = 1, 2$$

where $[\psi_{ir}]_{i,r} = H_4$ and $[\omega_{jc}]_{j,c} = H_2$, and

$$\hat{\theta}_{rc} = \sum_{j=1}^{2} \omega_{jc} \left(\sum_{i=1}^{4} \psi_{ir} P_{i(j)} \right) \quad \text{for } r = 2,3,4;\ c = 1,2.$$

In this example we look at the $\hat{\theta}_{r1}$ for $r = 2,3,4$ and the $\hat{\theta}_{r2}$ for $r = 2,3,4$ separately.

(1) Firstly, for $r = 2,3,4$

$$\hat{\theta}_{r1} = \frac{1}{\sqrt{2}} \sum_{i=1}^{4} \psi_{ir} P_{i+}$$

whose interaction matrix is

$$\underline{\psi}_r \otimes \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right).$$

These are the parameters we would get if we collapsed the original cross-classification over *species*. The three parameters $\hat{\theta}_{21}, \hat{\theta}_{22}$ and $\hat{\theta}_{23}$ can be interpreted as follows:

$\hat{\theta}_{21}$: *diameter* main-effect parameter

$\hat{\theta}_{31}$: *height* main-effect parameter

$\hat{\theta}_{41}$: *diameter* $\star$ *height* interaction parameter.

(2) Secondly, for $r = 2,3,4$

$$\hat{\theta}_{r2} = \frac{1}{\sqrt{2}} \sum_{i=1}^{4} (P_{i(1)} - P_{i(2)})$$

whose interaction matrix is

$$\underline{\psi}_r \otimes \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right).$$

These three parameters now contrast the two species categories on top of the $\underline{\psi}_r$ *diameter* $\star$ *height* contrast. The interpretations are now:

$\hat{\theta}_{22}$: *diameter* $\star$ *species* interaction parameter

$\hat{\theta}_{32}$: *height* $\star$ *species* interaction parameter

$\hat{\theta}_{42}$: three factor, *diameter* $\star$ *height* $\star$ *species* parameter.

The parameter table

	$\underline{\omega}_1$	$\underline{\omega}_2$
$\underline{\psi}_2$	188.4	122.0
	(33.6)	(33.6)
$\underline{\psi}_3$	-227.3	-109.0
	(33.0)	(33.0)
$\underline{\psi}_4$	-40.1	-89.2
	(35.4)	(35.4)

The criterion table

	$\underline{\omega}_1$	$\underline{\omega}_2$
$\underline{\psi}_2$	-33.2	-12.6
$\underline{\psi}_3$	-49.5	-9.7
$\underline{\psi}_4$	0.9	-5.5

The selection criterion gives that only $\hat{\theta}_{41}$, the *diameter* $\star$ *height* interaction parameter should be excluded. One may feel disinclined to exclude $\hat{\theta}_{41}$ while including $\hat{\theta}_{42}$, the three-factor *species* $\star$ *diameter* $\star$ *height* interaction parameter (c.f. Section 5.4 and the hierarchy principle). If so one can either include both $\hat{\theta}_{41}$ and $\hat{\theta}_{42}$, which gives the saturated model, or exclude both. The model corresponding to the second of these options has, in fact, the lower estimated expected discrepancy.

For the time being however, we fit the model with only $\hat{\theta}_{41}$ excluded. The fitted probabilities are shown.

perch height (in feet)	perch diameter (in inches)	species: sagrei	species: distichus
> 4.75	≤ 4.0	20.9	26.3
	> 4.0	5.3	15.3
≤ 4.75	≤ 4.0	51.0	28.4
	> 4.0	22.8	30.0
		100.0	100.0

THE CAMP DATA. For this data set there are three explanatory variables *race, origin* and *location.* There is one response variable, *preference*, with a quite complicated structure among its categories. The matrix of sample proportions, expressed as percentages, is shown.

race		Black				White			
origin		North		South		North		South	
location		North	South	North	South	North	South	North	Sou
preference									
prefer to stay		38.0	6.0	30.0	34.0	25.0	19.0	15.0	42
prefer to move	North	37.0	63.0	14.0	14.0	40.0	48.0	13.9	8
	South	7.0	12.0	31.0	29.0	11.0	9.0	48.9	34
	Undecided	7.9	11.0	13.0	13.0	13.0	15.0	11.1	8
undecided		10.1	8.0	12.1	10.0	11.0	9.0	11.1	8
Totals		100	100	100	100	100	100	100	10

It is convenient to regard the explanatory variable *location* $\times$ *origin* $\times$ *race* cross-classification as a single classification. For this classification the model basis that we will use is obtained from $H_8 = H_2 \otimes H_2 \otimes H_2$ by a simple re-ordering of the columns. The (non-normalised) basis is shown below. The letters L, 0 and R stand for *location, origin* and *race* respectively.

race	origin	location		L	O	R	L * O	L * R	O * R	L * O * R
		N	1	1	1	1	1	1	1	1
	N									
		S	1	-1	1	1	-1	-1	1	-1
Black										
		N	1	1	-1	1	-1	1	-1	-1
	S									
		S	1	-1	-1	1	1	-1	-1	1
		N	1	1	1	-1	1	-1	-1	-1
	N									
		S	1	-1	1	-1	-1	1	-1	1
White										
		N	1	1	-1	-1	-1	-1	1	1
	S									
		S	1	-1	-1	-1	1	1	1	-1

As regards *preference*, two possible bases were presented in Section 3.2

(A) The first of the two bases is

prefer to stay		1	1	0	1	0
prefer	North	1	1	0	-1/2	1
to	South	1	1	0	-1/2	-1
move	undecided	1	-3/2	1	0	0
undecided		1	-3/2	-1	0	0

whose columns will be denoted by $(\underline{\psi}_1, \ldots, \underline{\psi}_5)$. Roughly speaking the vectors $\underline{\psi}_2, \ldots, \underline{\psi}_5$ respectively contrast

- the "decideds" with the "undecideds"
- between the undecideds
- among the decideds, those who prefer to stay with those who do not
- among those who want to move, North versus South.

The parameter table

	·	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_2$	480.3	-23.8	3.7	3.6	30.0	17.0	67.4	-10.6
	(12.5)	(12.5)	(12.5)	(12.5)	(12.5)	(12.5)	(12.5)	(12.5)
$\underline{\psi}_3$	31.9	-28.1	12.4	-7.9	-17.8	-8.0	-27.4	2.3
	(11.0)	(11.0)	(11.0)	(11.0)	(11.0)	(11.0)	(11.0)	(11.0)
$\underline{\psi}_4$	2.8	-40.4	145.1	-28.1	-285.8	-205.0	60.0	-17.7
	(16.5)	(16.5)	(16.5)	(16.5)	(16.5)	(16.5)	(16.5)	(16.5)
$\underline{\psi}_5$	140.0	-104.8	605.1	105.4	-50.1	-10.1	-40.1	-44.7
	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)

The criterion table

	·	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\underline{\psi}_2$	-230.3	-0.3	-0.3	0.3	-0.6	0.0	-4.2	0.2
$\underline{\psi}_3$	-0.8	-0.5	0.1	0.2	-0.1	0.2	-0.5	0.2
$\underline{\psi}_4$	0.5	-1.1	-20.5	-0.2	-81.1	-41.5	-3.1	0.2
$\underline{\psi}_5$	-19.1	-10.5	-365.7	-10.6	-2.0	0.4	-1.1	-1.5

There are many parameters which are to be included in the fitted model. This is partly due to the large sample size which allows more parameters to be estimated reliably. (A total of 10289 recruits were questioned.)

There are too many parameters for us to comment on each individually and we will restrict attention to those parameters which make the largest reduction to the estimated expected discrepancy.

(1) The largest reduction comes from $\hat{\theta}_{53}$ – the "*origin* ⋆ (North or South)" interaction parameter – and indicates that *origin* played an important role in recruits deciding whether they wanted to move to a camp in the North or the South. The table below shows the overall effect of the incorporation of this parameter into the model, which indicates that there is a tendency for recruits to indicate that they

would like to move to a camp in the region from which they originate.

		origin	
		North	South
prefer to move	North	+	−
	South	−	+

(2) The next largest reduction comes from $\hat{\theta}_{21}$ – the "decided or undecided" main effect parameter. The importance of this parameter is simply due to the markedly smaller proportions in the two undecided rows than in the other rows.

(3) The next two largest reductions come from $\hat{\theta}_{45}$ and $\hat{\theta}_{46}$. The common subscript 4, refers to the contrast between those who prefer to stay and those who have a definite preference for where they want to move to. The inclusion of $\hat{\theta}_{45}$ indicates that there is interaction between this effect and the *location* $\star$ *origin* interaction. More specifically, those recruits whose location and origin values are the same are more likely to prefer to stay where they are. (The relative importance of this parameter and of $\hat{\theta}_{53}$ would suggest that the strongest tendency is for recruits to want to be in camps in their own region of origin.) As for $\hat{\theta}_{46}$; this models a tendency for Blacks in camps in the North and Whites in camps in the South to be more inclined to want to stay in their present camp.

The fitted conditional probabilities obtained from the model which contains only those parameters which assist in decreasing the estimated expected discrepancy, are shown.

race		Black				White			
origin		North		South		North		South	
location		North	South	North	South	North	South	North	Sou
preference									
prefer to stay		37.4	6.6	30.2	33.9	25.6	18.4	15.0	42
prefer	North	37.3	62.5	13.6	14.3	39.5	48.4	14.3	7
to	South	6.8	12.0	30.1	29.8	11.1	8.9	49.8	33
move	Undecided	8.2	10.8	14.3	12.7	12.2	14.7	10.3	8
undecided		10.3	8.3	11.8	9.2	11.6	9.5	10.6	8
Totals		100	100	100	100	100	100	100	10

(B) We now consider the second of the proposed bases for *preference.*

prefer to stay		1	1	-3	0	0
prefer	North	1	1	1	1	1
to	South	1	1	1	1	-1
move	Undecided	1	1	1	-2	0
undecided		1	-4	0	0	0

The parameter and criterion tables are shown below. A study of these reveals the same trends that were evident when the first basis was used. Furthermore in terms of the (total) estimated expected discrepancy there is very little to choose between the two fitted models (0,7954 for the first and 0,7950 for the second).

The parameter table

	·	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_2$	319.3	-36.8	12.0	-4.1	4.3	4.1	19.6	-4.7
	(11.4)	(11.4)	(11.4)	(11.4)	(11.4)	(11.4)	(11.4)	(11.4)
$\underline{\psi}_3$	-117.5	-37.5	138.3	-29.1	-281.0	-199.4	33.2	-13.4
	(16.0)	(16.0)	(16.0)	(16.0)	(16.0)	(16.0)	(16.0)	(16.0)
$\underline{\psi}_4$	340.4	-15.0	44.0	-2.1	-62.7	-51.1	86.0	-15.1
	(12.7)	(12.7)	(12.7)	(12.7)	(12.7)	(12.7)	(12.7)	(12.7)
$\underline{\psi}_5$	139.9	-104.8	605.1	105.4	-50.1	-10.1	-40.1	-44.7
	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)	(15.6)

The criterion table

	·	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_2$	-101.7	-1.1	0.1	0.2	0.2	0.2	-0.1	0.2
$\underline{\psi}_3$	-13.3	-0.9	18.6	-0.3	-78.4	-39.3	-0.6	0.3
$\underline{\psi}_4$	-115.6	0.1	-1.6	0.3	-3.6	-2.3	-7.1	0.1
$\underline{\psi}_5$	-19.1	-10.5	-365.7	-10.6	-2.0	0.4	-1.1	-1.5

THE BEETLE DATA

This data concerns the toxicity of an insecticide to a beetle species. The insecticide was administered at six dosage levels to six groups of beetles and the proportion surviving observed. In Section 3.2 two orthogonal polynomial bases were proposed for use in connection with the dose variable. The first of these was obtained by applying the Gram-Schmidt procedure to the actual dose values, while the second is the standard orthogonal polynomial basis obtained by using six equally spaced values.

Using the first basis, the selection criterion indicates that only the constant and the linear parameter are necessary, with a total estimated expected discrepancy of -.148. Using the second basis the constant, linear and quadratic terms are included, (the quadratic term only just). The total estimated expected discrepancy is slightly higher at -.139.

The mortality rates (as percentages)

	dose					
	12.08	14.49	16.31	18.31	20.44	22.36
sample	40.00	57.14	66.00	60.00	66.00	67.35
model, first basis	46.06	53.24	57.91	62.09	66.82	70.36
model, second basis	42.80	53.75	61.47	65.97	67.23	65.27

The parameter and criterion table, first basis

	·	x	x^2	x^3	x^4	x^5
parameter	326.1	283.0	-127.4	66.6	11.3	-74.3
std. deviation	97.7	97.8	97.6	98.6	97.2	97.4
contribution	-87.3	-61.0	2.8	15.0	18.8	13.5

The parameter and criterion table, second basis

	·	x	x^2	x^3	x^4	x^5
parameter	326.1	265.9	-139.5	104.1	-26.9	-68.5
std. deviation	97.7	97.6	97.4	97.9	98.0	97.6
contribution	-87.3	-51.6	-0.5	8.3	18.5	14.3

THE ESKIMO DATA

This data involves the *incidence* of *torus mandibularis* by age (six categories) for three Eskimo populations. The sample proportions, expressed as percentages, are:

		population			
incidence	age	Igloolik	Hall Beach	Aleut	Average
present	1-10	1.6	2.3	6.5	3.5
	11-20	6.0	6.3	2.8	5.0
	21-30	10.2	7.0	5.6	7.6
	31-40	9.8	7.8	8.3	8.6
	41-50	5.1	6.3	5.6	5.7
	50+	7.0	4.7	6.5	6.1
absent	1-10	27.3	30.5	14.8	24.2
	11-20	15.6	20.3	18.5	18.1
	21-30	12.1	9.4	13.9	11.8
	31-40	3.2	3.1	7.4	4.6
	41-50	1.3	1.6	6.5	3.1
	50+	1.0	0.8	3.7	1.8

We begin the analysis of this data set by giving the model bases that will be used for each of the variables.

For *incidence* we use the model basis

$$H_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

whose columns will be denoted by

$$(\underline{\phi}_1, \underline{\phi}_2).$$

For *age* we use the standard orthonormal polynomial basis of order 6×6, (see Section 3.2), whose columns will be denoted by

$$(\underline{\psi}_0, \underline{\psi}_1, \ldots, \underline{\psi}_5)$$

where the columns are labelled in this way so that $\underline{\psi}_n$ refers to a polynomial of order n.

For *population* we use the model basis obtained by normalising

$$\begin{array}{l} \text{Igloolik} \\ \text{Hall Beach} \\ \text{Aleut} \end{array} \begin{pmatrix} 1 & 0.5 & 1 \\ 1 & 0.5 & -1 \\ 1 & -1.0 & 0 \end{pmatrix}$$

whose columns will be denoted by

$$(\underline{\omega}_1, \underline{\omega}_2, \underline{\omega}_3).$$

The parameter associated with $\underline{\phi}_i, \underline{\psi}_n$ and $\underline{\omega}_p$ will be denoted by $\hat{\theta}_{inp}$ for $i = 1,2;\ n = 0,\ldots,5;\ p = 1,2,3$.

The parameter and criterion tables are:

The parameter table

		$\underline{\omega}_1$	$\underline{\omega}_2$	$\underline{\omega}_3$
		-217.2	-63.5	23.2
	$\underline{\psi}_1$	(21.8)	(24.2)	(19.0)
		20.6	30.3	-17.7
	$\underline{\psi}_2$	(23.9)	(25.9)	(21.7)
		24.1	1.5	8.3
$\underline{\phi}_1$	$\underline{\psi}_3$	(23.8)	(25.9)	(21.4)
		11.7	5.6	31.8
	$\underline{\psi}_4$	(23.6)	(25.9)	(21.0)
		-6.9	-3.9	-15.3
	$\underline{\psi}_5$	(23.4)	(26.1)	(20.4)
		-135.9	8.7	21.7
	$\underline{\psi}_0$	(22.8)	(24.8)	(20.6)
		263.5	73.2	-9.6
	$\underline{\psi}_1$	(20.8)	(23.5)	(17.7)
		-95.5	-85.0	5.1
	$\underline{\psi}_2$	(23.3)	(25.6)	(20.8)
		-16.2	43.5	11.3
$\underline{\phi}_2$	$\underline{\psi}_3$	(23.7)	(25.7)	(21.5)
		34.9	-23.8	-1.5
	$\underline{\psi}_4$	(23.5)	(25.8)	(21.0)
		22.9	0.0	13.3
	$\underline{\psi}_5$	(23.4)	(26.1)	(20.4)

The criterion table

		$\underline{\omega}_1$	$\underline{\omega}_2$	$\underline{\omega}_3$
	$\underline{\psi}_1$	-46.2	-2.9	0.2
	$\underline{\psi}_2$	0.7	0.4	0.6
$\underline{\phi}_1$	$\underline{\psi}_3$	0.6	1.3	0.9
	$\underline{\psi}_4$	1.0	1.3	-0.1
	$\underline{\psi}_5$	1.1	1.3	0.6
	$\underline{\psi}_0$	-17.4	1.2	0.4
	$\underline{\psi}_1$	-68.6	-4.3	0.5
	$\underline{\psi}_2$	-8.0	-5.9	0.8
$\underline{\phi}_2$	$\underline{\psi}_3$	0.9	-0.6	0.8
	$\underline{\psi}_4$	-0.1	0.8	0.9
	$\underline{\psi}_5$	0.6	1.4	0.7

Consider first the upper half of the criterion table, which is the half in which *incidence* is ignored. A remarkable feature is that this half of the table contains very few negative entries. This demonstrates the age distribution in the three populations is readily modelled with the standard orthogonal polynomial basis. The two most important parameters in this half, $\hat{\theta}_{111}$ and $\hat{\theta}_{112}$; both model a linear trend in age. The first relates to an overall trend in all populations, while the second contrasts the linear trend between the first two and the last population. The only other parameter in this half which makes a negative contribution is $\hat{\theta}_{143}$ which involves interaction between

- a fourth degree polynomial trend in age, and
- the contrast between the first two populations.

It seems unlikely that this indicated difference involving such a high order polynomial is a real feature of the operating model, especially in view of the small magnitude of the contribution, and is more likely to have arisen as a result of sam-

pling variation. Consequently one would probably not include this parameter in the final model.

Thus, when *incidence* is ignored we are left with two parameters which model the probabilities as having a linear trend with *age*; one trend for the first two populations and a different trend for the last.

We consider next the bottom half of the table. The inclusion of any parameters from this half indicates a difference of some sort between the proportions with and without *torus mandibularis*. The following observations can be made.

1. None of the contributions in the third column in this half ~~are~~ is negative which indicates that it is not considered worthwhile incorporating any parameters which cause the fitted probabilities to differ between the first two populations. This suggests that there is no difference in the incidence of *torus mandibularis* by age between these two populations.

2. The contribution of $\hat{\theta}_{211}$ makes the largest reduction, indicating that there is a distinct difference as regards linear trend for the two *incidence* categories.

3. The two parameters $\hat{\theta}_{232}$ and $\hat{\theta}_{241}$, like $\hat{\theta}_{143}$ in the upper half, both make small negative contributions and involve high order polynomial terms, so that it may be best not to include them into the final fitted model.

The conditional probabilities obtained when including only those parameters whose contributions are less than -0.001 (i.e. excluding the three parameters mentioned above) are shown.

incidence	age	population		
		Igloolik	Hall Beach	Aleut
present	1-10	1.3	1.3	5.8
	11-20	6.0	6.0	5.5
	21-30	8.6	8.6	5.5
	31-40	9.1	9.1	5.8
	41-50	7.5	7.5	6.4
	50+	3.9	3.9	7.4
absent	1-10	28.1	28.1	17.1
	11-20	18.4	18.4	14.9
	21-30	10.6	10.6	12.4
	31-40	5.0	5.0	9.6
	41-50	1.5	1.5	6.5
	50+	0.0	0.0	3.0

THE VISION DATA. This data set is concerned with vision in the right and left eyes. The sample proportions expressed as percentages:

grade of right eye	grade of left eye: highest (1)	second (2)	third (3)	lowest (4)	Totals
highest (1)	25.32	3.45	2.62	1.08	32.47
second (2)	3.58	15.24	4.47	0.83	24.12
third (3)	2.22	4.66	17.98	2.68	27.54
lowest (4)	1.83	1.05	3.27	10.21	16.36
Totals	32.95	24.40	28.34	14.80	100

Using the notation defined when the data set was introduced in 3.2, a suitable basis for

$$\begin{pmatrix} U \\ D \\ L \end{pmatrix}$$

is the partitioned matrix

$$\begin{pmatrix} \underline{1}_6 & 0 & \underline{1}_6 & \Phi & \underline{1}_6 & \Phi \\ \underline{1}_4 & \Omega & -3\underline{1}_4 & 0 & 0 & 0 \\ \underline{1}_6 & 0 & \underline{1}_6 & \Phi & -\underline{1}_6 & -\Phi \end{pmatrix}$$

where Ω contains contrasts for the four diagonal cells

$$\begin{matrix} (1,1) \\ (2,2) \\ (3,3) \\ (4,4) \end{matrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \\ -3 & 0 & 0 \end{pmatrix}$$

and Φ contains contrasts for either the upper or lower off-diagonal halves:

$$\begin{array}{cc} (1,2) & (2,1) \\ (1,3) & (3,1) \\ (2,3) & (3,2) \\ (1,4) & (4,1) \\ (2,4) & (4,2) \\ (3,4) & (4,3) \end{array} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & -1/2 & 1 & 0 & 0 \\ 1 & -1/2 & -1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 0 & -2 \\ -1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

The first two vectors in this matrix are used to contrast the row (or column) totals, while the next three contrast cells within individual rows (or columns).

Let $\Psi = (\underline{\psi}_1, \ldots, \underline{\psi}_{16})$ denote the columns of the joint basis and let π_i for $i = 1, \ldots, 16$ denote the (operating model) probability with which a randomly selected patient falls into the ith cell of

$$\begin{pmatrix} U \\ D \\ L \end{pmatrix}.$$

The saturated model can then be written as

$$M_i(\hat{\theta}) = \frac{1}{16} + \sum_{q=2}^{16} \phi_{iq}\hat{\theta}_q \quad \text{for } i = 1, \ldots, 16$$

where

$$\hat{\theta}_q = \Sigma_i \phi_{iq} P_i \quad \text{for } q = 2, \ldots, 16.$$

For purposes of interpretation the parameters can be represented in the table shown below.

	θ_2
main diagonal	θ_3
	θ_4
diagonal versus off-diagonal	θ_5

	upper and lower off-diagonal halves	
	averaged over	contrasted
average	–	θ_{11}
row or	θ_6	θ_{12}
column totals	θ_7	θ_{13}
within individual	θ_8	θ_{14}
rows or columns	θ_9	θ_{15}
	θ_{10}	θ_{16}

Remarks.

1. If $\theta_{11},\ldots,\theta_{16}$ are excluded from any model then the corresponding cells in the upper and lower off-diagonal halves will be identical; which means that the modelled cell probabilities will be symmetrical. Clearly the ability of the modelling procedure to provide a *symmetric model* is a highly desirable property, both for this particular data set and others like it.

2. The modelling procedure is also able to produce models which are not symmetric but whose corresponding row and column totals are equal, (i.e. a so-called *model of marginal homogeneity*). Such a model will not contain $\hat{\theta}_{11},\hat{\theta}_{12}$ and $\hat{\theta}_{13}$ although it may contain some of $\hat{\theta}_{14},\hat{\theta}_{15}$ and $\hat{\theta}_{16}$.

The parameter table

	80.58	(1.75)
main diagonal	18.76	(3.10)
	71.32	(5.52)
diagonal versus off-diagonal	-252.61	(0.68)

	upper and lower off-diagonal halves averaged over		contrasted	
average	–		-2.76	(0.82)
row or	31.08	(0.80)	3.29	(0.82)
column totals	0.27	(0.95)	-1.34	(0.95)
within individual	-21.44	(1.63)	2.93	(1.64)
rows or columns	17.74	(1.26)	-1.70	(1.24)
	13.27	(0.58)	-1.16	(0.58)

The criterion table

	-64.2
main diagonal	-2.4
	-49.6
diagonal versus off-diagonal	-637.7

	upper and lower off-diagonal halves averaged over	contrasted
average	–	0.1
row or	-9.5	0.1
column totals	0.2	0.2
within individual	-4.4	0.1
rows or columns	-3.0	0.1
	-1.7	0.1

The selection criterion indicates that none of the parameters $\hat{\theta}_{11}, \ldots, \hat{\theta}_{16}$ should be included in the model. This means that the modelled cell probabilities are symmetric. The fitted probabilities, expressed as percentages, are given below.

grade of right eye	grade of left eye highest (1)	second (2)	third (3)	lowest (4)	Totals
highest (1)	25.32	3.50	2.43	1.20	32.45
second (2)	3.50	15.24	4.57	0.94	24.25
third (3)	2.43	4.57	17.98	2.98	27.96
lowest (4)	1.20	0.94	2.98	10.21	15.33
Totals	32.45	24.25	27.96	15.33	100

CHAPTER 5
LOGLINEAR MODELS

In this chapter we consider loglinear models, that is, models in which the logs of the cell probabilities are expressed as linear combinations of parameters. The construction of model bases for loglinear models can be carried out in the same way as it is for the linear case. However there are three areas in which linear and loglinear models do differ substantially.

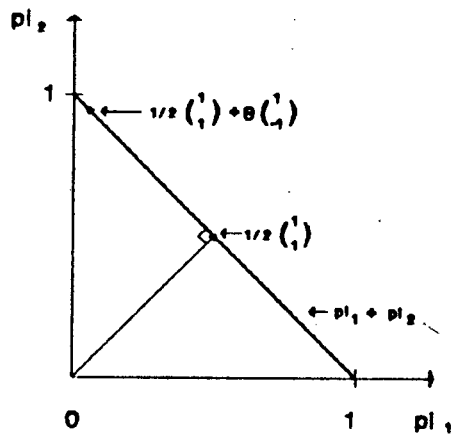
(1) In the linear case one compares the *difference* between probabilities, while in the loglinear case one is essentially comparing the *ratio* of probabilities.

(2) The second difference concerns the range of values which the modelled cell probabilities assume. In the loglinear case the fitted probabilities which are, for example in the univariate case, of the form

$$M_i(\hat{\theta}) = \exp(\sum_{q \in Q} \phi_{iq} \hat{\theta}_q) \quad \text{for } i = 1, \ldots, L$$

are always positive. In the linear case this property is not guaranteed.

(3) A third difference concerns the orthogonality of the parameters. For illustrative purposes consider a simple two-cell classification. That the two probabilities π_1 and π_2, must sum to one means that they must lie on a one-dimensional linear subspace of $\Re^2$.

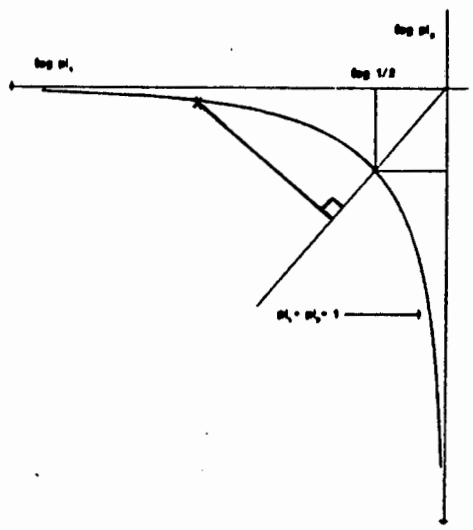


One may then take the point $\frac{1}{2}\binom{1}{1}$ as the starting point and any point on the subspace can be obtained as

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}+\theta\begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{pmatrix}$$

where, of course, $\binom{1/\sqrt{2}}{1/\sqrt{2}}$ and $\binom{1/\sqrt{2}}{-1/\sqrt{2}}$ are orthogonal to one another.

Now in "log space":



the subspace defined by $\pi_1+\pi_2=1$ is, in this space, not linear. It can be seen that, although any point in the subspace can be obtained as

$$\theta_1\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}+\theta_2\begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{pmatrix},$$

both θ_1 and θ_2 will vary for each different point in the space. Thus parameter orthogonality and all of its consequences in simplifying the model selection procedure that we had in the linear class of models, will no longer hold. Consequently the selection of loglinear models requires more computational effort than was needed for linear models.

In this chapter, basis loglinear models are introduced in Sections 1, 2 and 3. A particular class of models, to which selection is restricted, is introduced in Section 4, while in Section 5 basis loglinear models are fitted to each of the data sets introduced in Section 3.2.

5.1 SIMPLE CLASSIFICATIONS

We begin by looking at simple classifications although much of what follows here applies directly to multiway cross-classifications. In particular the estimation and selection procedures derived below are applicable, with only minor modification, to multivariate cases. Subsequent sections which deal with the multivariate cases will concentrate on the interpretation of the models, and in particular on the various forms of independence which can arise in a multiway cross-classification.

Consider then a univariate (multinomial) operating model with probabilities $\underline{\pi} = [\pi_i]_{i=1,\ldots L}$ and let $[\phi_{iq}]_{i=1,\ldots,L;\ q=1,\ldots,L}$ be a model basis for $\Re^L$. The class of loglinear approximating families is given by

$$\left\{\mathbf{M}(Q) \ : \ Q \subseteq \{1,\ldots,L\}, 1 \in Q\right\} \tag{1}$$

where

$$\mathbf{M}(Q) = \left\{\underline{M}(\theta) \ : \ \log\ M_i(\theta) = \sum_{q \in Q} \phi_{iq}\theta_q; \ \sum_i M_i(\theta) = 1\right\}.$$

For a model $\underline{M}(\theta) \in \mathbf{M}(Q)$ the essential part of the Kullback-Leibler discrepancy function is given by

$$\Delta(\underline{\pi}, \underline{M}(\theta)) = -\sum_i n_+ \pi_i \log\ M_i(\theta)$$

where n_+ is the sample size.

The minimum discrepancy parameters for an approximating family $\mathbf{M}(Q)$ are defined by

$$\theta^0(Q) = \arg\{\min\ \Delta(\underline{\pi}, \underline{M}(\theta)) \ : \ \underline{M}(\theta) \in \mathbf{M}(Q)\}.$$

The solution to this minimisation problem is found using the method of Lagrangian multipliers.

Theorem 1. The minimum discrepancy parameters for an approximating family $\mathbf{M}(Q)$ from the class given in (1) are the solutions to

$$\sum_i \phi_{iq}[\exp(\sum_{q \in Q} \phi_{iq}\theta_q^0(Q)) - \pi_i] = 0 \quad \text{for } q \in Q. \tag{2}$$

Proof. Define

$$G(\theta,\lambda) = -\sum_i n_+\pi_i \log M_i(\theta) + \lambda(\sum_i M_i(\theta) - 1)$$
$$= -\sum_i n_+\pi_i(\sum_{q\in Q}\phi_{iq}\theta_q) + \lambda(\sum_i \exp(\sum_{q\in Q}\phi_{iq}\theta_q) - 1)$$

where λ is a Lagrangian multiplier.

The $\theta_q^0(Q)$ are then found by simultaneously solving the equations

$$\left.\begin{aligned}\frac{\partial G(\theta,\lambda)}{\partial\theta_q} &= 0 \quad \text{for} \quad q\in Q\\ \frac{\partial G(\theta,\lambda)}{\partial\lambda} &= 0.\end{aligned}\right\}$$

Now

$$\frac{\partial G(\theta,\lambda)}{\partial\theta_1} = 0 \quad \text{iff} \quad \frac{1}{\sqrt{L}}(-n_+ + \lambda\Sigma_i M_i(\theta)) = 0$$
$$\frac{\partial G(\theta,\lambda)}{\partial\lambda} = 0 \quad \text{iff} \quad \Sigma_i M_i(\theta) = 1.$$

Solving these two equations simultaneously gives $\lambda = n_+$ and renders the second equation redundant.

The resulting system is

$$-n_+\sum_i \phi_{iq}\pi_i + n_+\sum_i \phi_{iq}\exp(\sum_{q\in Q}\phi_{iq}\theta_q^0(Q)) = 0 \quad \text{for } q\in Q,$$

which simplifies to

$$\sum_i \phi_{iq}[\exp(\sum_{q\in Q}\theta_q^0(Q)\phi_{iq}) - \pi_i] = 0 \quad \text{for } q\in Q. \quad \bullet$$

The system (2) is non-linear in the $\theta_q^0(Q)$ and, in general, closed form expressions for the minimum discrepancy parameters cannot be obtained. The two exceptions (for the univariate case) are:

(i) the saturated approximating family, in which case

$$\theta_q^0(\{1,\ldots,L\}) = \Sigma_i \phi_{iq} \log \pi_i \quad \text{for } q = 1,\ldots,L,$$

(ii) the family of models which contains the single parameter, θ_1 , in which case

$$\theta_1^0(\{1\}) = \sqrt{L} \log \frac{1}{L},$$

so that

$$M_i\Big(\theta_1^0(\{1\})\Big) = \frac{1}{L} \quad \text{for all } i.$$

In order to estimate the parameters one replaces each π_i in (2) by its sample analog P_i . The resulting system can be solved numerically by the well-known Newton-Raphson method. (See Appendix A.) The estimates thus obtained are denoted by $\hat{\theta}_q(Q)$, so that the fitted model from the family $\mathbf{M}(Q)$ is

$$M_i(\hat{\theta}(Q)) = \exp(\sum_{q \in Q} \phi_{iq} \hat{\theta}_q(Q)) \quad \text{for } i = 1,\ldots,L. \tag{3}$$

The interpretation of the parameters in these models is similar to that of the parameters in linear models. The only difference being that whereas in a linear model the inclusion of $\hat{\theta}_q(Q)$ involves the addition of $\underline{\phi}_q \hat{\theta}_q(Q)$ to the modelled cell probabilities, in a loglinear model $\underline{\phi}_q \hat{\theta}_q(Q)$ is added to the log of the modelled cell probabilities.

The expected discrepancy of the fitted model (3) from the approximating family $\mathbf{M}(Q)$ is

$$\begin{aligned} & - E_{\underline{\pi}}(\sum_i n_+ \pi_i (\sum_{q \in Q} \phi_{iq} \hat{\theta}_q(Q))) \\ = & - \sum_i n_+ \pi_i (\sum_{q \in Q} \phi_{iq} E_{\underline{\pi}}(\hat{\theta}_q(Q))). \end{aligned} \tag{4}$$

To evaluate $E_{\underline{\pi}}(\hat{\theta}_q(Q))$, particularly when the $\hat{\theta}_q(Q)$ are defined only implicitly as the solution to (1), is not easy. Having done this, one would then have to find

an estimator, preferably unbiased, for the whole of (4), which involves estimating terms of the type

$$\pi_i E_{\underline{\pi}}(\hat{\theta}_q(Q))$$

which, again, is not easy.

The cross-validated discrepancy. An approach which circumvents the above difficulties has been suggested by Linhart and Zucchini (1986b). The idea is to estimate the expected discrepancy directly, without having to first evaluate it. This can be achieved by making use of a one-item-out cross-validatory procedure which we will now outline.

Let $P(n_1,\ldots,n_L;\ \underline{\pi}, n_+)$ denote the probability of the sample cell counts under the operating model, i.e.

$$P(n_1,\ldots,n_L;\ \underline{\pi}, n_+) = \binom{n_+}{n_1 \ldots n_L} \pi_1^{n_1} \ldots \pi_L^{n_L}.$$

Now suppose that one observation from the sample is chosen at random and "hidden". Denote the resulting cell counts by $n_1^*,\ldots,n_L^*;\ \Sigma_i n_i^* = n_+^* = n_+ - 1$. From the family $\mathbf{M}(Q)$, let

$$\left[M_i(\hat{\theta}(Q);\ n_1^*,\ldots,n_L^*)\right]_{i=1,\ldots,L} \tag{5}$$

be the fitted model obtained using the reduced sample.

Theorem. An unbiased estimator of the expected discrepancy for the fitted model (5) is

$$-\frac{n_+ - 1}{n_+}\Sigma_i n_i \log M_i(\hat{\theta}(Q);\ n_1,\ldots,n_i - 1,\ldots,n_L) \tag{6}$$

where $n_i \log M_i(\hat{\theta}(Q);\ n_1,\ldots,n_i - 1,\ldots,n_L)$ is defined to be zero if $n_i = 0$.

Proof. The expected discrepancy for (5) is

$$\sum_{\substack{n_1^*,\ldots,n_L^* \geq 0 \\ \Sigma_i n_i^* = n_+^*}} \Delta(\underline{\pi}; M(\hat{\theta}(Q);\ n_1^*,\ldots,n_L^*))\ \ P(n_1^*,\ldots,n_L^*;\ \underline{\pi}, n_+^*)$$

which can be written as $\Sigma_i A_i$, where

$$A_i = -n_+^* \pi_i \sum_{\substack{n_1^*,\ldots,n_L^* \geq 0 \\ \Sigma_i n_i^* = n_+^*}} \log M_i(\hat{\theta}(Q);\, n_1^*,\ldots,n_L^*)\, P(n_1^*,\ldots,n_L^*;\, \underline{\pi}, n_+^*)\,.$$

Now put

$$B_i = -\frac{n_+ - 1}{n_+} \sum_{\substack{n_1,\ldots,n_L \geq 0 \\ \Sigma_i n_i = n}} n_i \log M_i(\hat{\theta}(Q);\, n_1,\ldots,n_i-1,,\ldots n_L)\, P(n_1,\ldots,n_L;\, \underline{\pi}, n_+)$$

where $n_i \log M_i(\hat{\theta}(Q);\, n_1,\ldots,n_i-1,\ldots,n_L)$ is defined to be zero if $n_i = 0$.

Then for each i

$$B_i = -\frac{n_+ - 1}{n_+} \sum_{\substack{n_1,\ldots,n_i-1,\ldots,n_L \geq 0 \\ (\Sigma_{j\neq i} n_i)+(n_i-1)=n-1}} \log M_i(\hat{\theta}(Q);\, n_1,\ldots,n_i-1,\ldots,n_L)\cdot$$

$$\frac{n_+!}{n_1!\ldots(n_i-1)!\ldots n_L!}\pi_1^{n_1}\ldots\pi_L^{n_L}$$

$$= -(n_+ - 1)\pi_i \sum_{\substack{n_1,\ldots,n_i-1,\ldots n_L \geq 0 \\ (\Sigma_{j\neq i} n_i)+(n_i-1)=n-1}} \log M_i(\hat{\theta}(Q);\, n_1,\ldots,n_i-1,\ldots,n_L)\cdot$$

$$P(n_1,\ldots,n_i-1,\ldots,n_L;\, \underline{\pi}, n_+ - 1)$$

$$= A_i\,.$$

Hence the expected discrepancy of (5) can be written as

$$\sum_i B_i = \sum_{\substack{n_1,\ldots,n_L \geq 0 \\ \Sigma_i n_i = n}} \left(-\frac{(n_+ - 1)}{n_+} \sum_i n_i \log M_i(\hat{\theta}(Q); n_1,\ldots,n_i-1,\ldots,n_L) \right).$$

$$P(n_1,\ldots,n_L;\, \underline{\pi}, n_+)$$

which is the expectation of (6) under the operating model.Thus (6) is the unbiased estimator of the expected discrepancy for (5). •

The expression given in (6) is called the *cross-validated discrepancy.* The fitted model with the smallest cross-validated discrepancy is estimated to be optimal for the reduced data set. In practice we do not actually discard one observation but

simply use (6) as the criterion. In effect this introduces a bias and we are assuming that this will have little effect on the ranking of the approximating models in terms of their expected discrepancies.

The cross-validated expected discrepancy can be written in orthogonal form with each parameter contributing separately to the total. This is done by writing (6) as

$$-\frac{n_+ - 1}{n_1}\sum_i n_i\Big(\sum_{q\in Q}\phi_{iq}\hat{\theta}_q(Q;\ n_1,\ldots,n_i-1,\ldots,n_L)\Big)$$
$$=\frac{n_+ - 1}{n_+}\sum_{q\in Q}\Big[-\sum_i \phi_{iq}n_i\hat{\theta}_q(Q;\ n_1,\ldots,n_i-1,\ldots,n_L)\Big]$$

where now $n_i\hat{\theta}_q(Q;\ n_i,..,n_i-1,\ldots,n_L)$ is defined to be zero if $n_i = 0$. The quantity within the square brackets will be referred to as the *contribution of* $\hat{\theta}_q(Q)$ to the cross-validated discrepancy, and denoted by $C(\hat{\theta}_q(Q))$.

Note that in order to evaluate the contributions of the $\hat{\theta}_q(Q)$ for a *given* approximating family one must:

(A) for each i $(i = 1,\ldots,L)$ reduce the ith cell count by one and re-estimate the parameters to compute

$$\hat{\theta}_q(Q;\ n_1,\ldots,n_i-1,\ldots,n_L)$$

for each $q \in Q$ (using numerical methods),

(B) for each $q \in Q$, compute

$$-\Sigma_i\phi_{iq}n_i\hat{\theta}_q(Q;\ n_1,\ldots,n_i-1,\ldots,n_L).$$

Since the parameters are not orthogonal each parameter does not make a fixed contribution to the cross-validated discrepancy. Thus in order to find the fitted model with the smallest cross-validated discrepancy from a class of approximating families, one has to actually fit each of the models and compute their cross-validated discrepancies.

For small classifications (involving up to about 6 cells) computations are manageable. For larger tables the computational effort required becomes excessive, and one has to adopt an heuristic approach.

One can, for example, examine the contribution to the criterion of each parameter in the *saturated* model. The parameters can then be (subjectively) partitioned into three sets; those which make a large negative contribution and are therefore likely to be present in the optimal model; those making a large positive contribution and which are likely to be absent in the optimal model; and the rest. We then only examine those approximating models which include the parameters in the first of these three sets and exclude those in the second.

Clearly, when using this procedure, we cannot be certain that the selected model is that which leads to the smallest criterion. In fact it is quite easy to construct artificial data sets for which the procedure would select a model which is far from optimal. In applying the procedure we are assuming that the contributions to the criterion of each parameter do not vary substantially across different approximating families. This assumption was justified in the case of the data sets to be discussed in Section 5.

In later sections a rule will be given which limits the number of "permissible" models and so makes the model selection process somewhat easier. •

5.2 TWO-WAY CROSS-CLASSIFICATIONS

For many purposes two-way classifications can be regarded as simple classifications, but there are also reasons for not doing so. Firstly it is convenient for the purposes of interpretations to explicitly regard the variables separately. For example, we might be particularly interested to investigate the joint behaviour of the variables, to see whether they should be modelled as being independent. Secondly there is a close link between two-way multinomial models and product-multinomial models which, in effect, allows us to use the same results and algorithms for both cases.

This section contains four subsections:

A: multinomial two-way tables

B: product-multinomial two-way tables

C: hierarchical models for two-way tables

D: standard hierarchical models for two-way tables.

The first subsection simply introduces the class of approximating families, the minimum discrepancy parameters and the bivariate version of the cross-validated discrepancy for multinomial two-way tables. The second does the same for the product-multinomial case, and in it we outline the nature of the relationship between the multinomial and product-multinomial cases. The final two subsections deal with hierarchical models and their special properties.

A. MULTINOMIAL TWO-WAY TABLES

Consider a $R \times C$ cross-classification with a multinomial operating model with cell probabilities

$$\pi = [\pi_{ij}]_{i=1,\ldots,R;\ j=1,\ldots,C}.$$

Let Ψ and Ω be model bases for Y and X respectively. The class of loglinear approximating families considered is

$$\{\mathbf{M}(Q) \ : \ Q \subseteq R \times C,\ (1,1) \in Q\}$$

where

$$\mathbf{M}(Q) = \left\{ [M_{ij}(\theta)]_{i,j} \; : \; \log M_{ij}(\theta) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc}; \; \sum_i \sum_j M_{ij}(\theta) = 1 \right\}.$$

The minimum discrepancy parameters for an approximating family $\mathbf{M}(Q)$ are the solutions to the system of equations

$$\sum_i \sum_j \psi_{ir}\omega_{jc} \left[\exp\Big(\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta^0_{rc}(Q) \Big) - \pi_{ij} \right] = 0 \quad \text{for } (r,c) \in Q.$$

These parameters are estimated by replacing each π_{ij} by its sample analog P_{ij}, and then solving the resulting system (using some numerical method). If $\hat{\theta}(Q)$ are the resulting estimates then the fitted model $[M_{ij}(\hat{\theta}(Q))]_{i,j}$, from the family $\mathbf{M}(Q)$, has

$$\log M_{ij}(\hat{\theta}(Q)) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\hat{\theta}_{rc} \quad \text{for all } i,j. \tag{1}$$

In these models the role played by each parameter, say $\hat{\theta}_{rc}$, is determined by its interaction matrix, $[\psi_{ir}\omega_{jc}]_{i=1,\ldots,R;\; j=1,\ldots,C}$, in that the interaction matrix determines the way in which the contributions due to $\hat{\theta}_{rc}$ are added to the table of the $\log M_{ij}(\hat{\theta}(Q))$.

The parameters can again be divided into four distinct groups (c.f. Section 3.1):

- $\hat{\theta}_{11}$, the constant parameter
- $\{\hat{\theta}_{r1}\}_{r\in\{2,\ldots,R\}}$, the row-effect parameters
- $\{\hat{\theta}_{1c}\}_{c\in\{2,\ldots,C\}}$, the column-effect parameters
- $\{\hat{\theta}_{rc}\}_{r\in\{2,\ldots,R\},c\in\{2,\ldots,C\}}$, the row $\star$ column interaction parameters.

The following result provides further insight into the role played by the different parameters.

Theorem 1. Consider a loglinear model for a two-way cross-classification from an approximating family $\mathbf{M}(Q)$.

(i) If $\{(1,c) : c = 1,\ldots,C\} \subseteq Q$ (i.e. the model contains all of the column effect parameters θ_{1c} for $c = 1,\ldots,C$), then

$$M_{+j}(\theta^0(Q)) = \pi_{+j} \quad \text{for } j = 1,\ldots,C$$
$$M_{+j}(\hat{\theta}(Q)) = P_{+j} \quad \text{for } j = 1,\ldots,C.$$

(ii) The same applies with row and column interchanged.

Proof. (i) The minimum discrepancy parameters are found by solving

$$\Sigma_i\Sigma_j\psi_{ir}\omega_{jc}[M_{ij}(\theta^0(Q)) - \pi_{ij}] = 0 \quad \text{for } (r,c) \in Q.$$

In particular then, since $(1,c) \in Q$ for $c = 1,\ldots,C$

$$\Sigma_i\Sigma_j \frac{1}{\sqrt{R}}\omega_{jc}[M_{ij}(\theta^0(Q)) - \pi_{ij}] = 0 \quad \text{for } c = 1,\ldots,C, \tag{2}$$

i.e.

$$\Sigma_j\omega_{jc}[M_{+j}(\theta^0(Q)) - \pi_{+j}] = 0 \quad \text{for } c = 1,\ldots,C$$

which is a linear system of full rank (the coefficient matrix is Ω which is orthogonal), so that the system has a unique solution, namely

$$M_{+j}(\theta^0(Q)) = \pi_{+j} \quad \text{for } j = 1,\ldots,C.$$

The parameters are estimated by replacing each π_{ij} by P_{ij} in (2), so that

$$M_{+j}(\hat{\theta}(Q)) = P_{+j} \quad \text{for } j = 1,\ldots,C.$$

(ii) The proof is similar to the above. •

The expected discrepancy of the fitted model (1) can be estimated by the bivariate version of the cross-validated discrepancy, namely

$$-\frac{n_{++}-1}{n_{++}}\Sigma_j\Sigma_i n_{ij} \log M_{ij}(\hat{\theta}(Q);\ N^{ij})$$

where

(1) $n_{ij} \log M_{ij}(\hat{\theta}(Q);\ N^{ij})$ is defined to be zero if $n_{ij} = 0$; and

(2) $(\hat{\theta}(Q); N^{ij})$ are the parameter estimates obtained when the count in the (i,j)th cell has been decreased by one.

B. PRODUCT-MULTINOMIAL TWO-WAY TABLES

Just as all multinomial cross-classifications can be reduced to one-way (multinomial) classifications, so all product-multinomial cross-classifications can be reduced to two-way (product-multinomial) cross-classifications. Hence we consider the two-way product-multinomial case in some detail now.

Consider a $R \times C$ product-multinomial operating model with indexed probabilities $\pi_{i(j)}$ where $\Sigma_i \pi_{i(j)} = 1$, for $j = 1, \ldots, C$. Clearly the modelled probabilities, $M_{i(j)}(\theta)$ must satisfy the same constraints. This leads us to consider approximating families

$$\mathbf{M}(Q) = \left\{ \left[M_{i(j)}(\theta)\right]_{i,j} \ :\ \log M_{i(j)}(\theta) = \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc}, \right.$$
$$\left. \sum_i M_{i(j)}(\theta) = 1 \quad \text{for all } j \right\}.$$

To distinguish between the essential and optional parameters we will sometimes write

$$\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta_{rc} = \frac{1}{\sqrt{R}} \sum_{c=1}^{C} \omega_{jc}\theta_{1c} + \sum_{(r,c)\in Q^*} \psi_{ir}\omega_{jc}\theta_{rc}$$

where

$$Q = Q^* \cup \{(1,c)\ :\ c = 1, \ldots, C\}.$$

For a model $M(\theta) \in \mathbf{M}(Q)$ the essential part of the Kullback-Leibler discrepancy function is given by

$$\Delta(\pi, M(\theta)) = -\Sigma_j \Sigma_i n_{+j} \pi_{i(j)} \log M_{i(j)}(\theta).$$

Minimum discrepancy parameters. The minimum discrepancy parameters for an approximating family $\mathbf{M}(Q)$ are defined by

$$\theta^0(Q) = \arg\{\min\ \Delta(\pi, M(\theta))\ :\ M(\theta) \in \mathbf{M}(Q)\}.$$

This minimisation problem involves C constraints $(\Sigma_i M_{i(j)}(\theta) = 1$ for $j = 1, \ldots, C)$ and is solved using C Lagrangian multipliers.

Theorem 2. The minimum discrepancy parameters for an approximating family $\mathbf{M}(Q)$, are the solutions to

$$\sum_j \sum_i \psi_{ir}\omega_{jc}\left[n_{+j}\exp\left(\sum_{(r,c)\in Q}\psi_{ir}\omega_{jc}\theta^0_{rc}(Q)\right) - n_{+j}\pi_{i(j)}\right] = 0 \quad \text{for } (r,c) \in Q. \tag{3}$$

Proof. Define

$$G(\theta, \lambda_1, \ldots, \lambda_C) = \Sigma_j \Sigma_i n_{+j}\pi_{i(j)} \log M_{i(j)}(\theta) + \Sigma_j \lambda_j (1 - \Sigma_i M_{i(j)}(\theta)).$$

Then

$$\frac{\partial G}{\partial \theta_{rc}} = -\Sigma_j \Sigma_i n_{+j}\pi_{i(j)}\psi_{ir}\omega_{jc} + \Sigma_j \lambda_j (\Sigma_i \psi_{ir}\omega_{jc} M_{i(j)}(\theta))$$

$$\frac{\partial G}{\partial \lambda_j} = 1 - \Sigma_i M_{i(j)}(\theta).$$

In particular

$$\frac{\partial G}{\partial \theta_{1c}} = \frac{1}{\sqrt{R}}\Sigma_j \omega_{jc} n_{+j} + \Sigma_j \lambda_j (\frac{1}{\sqrt{R}}\omega_{jc}\Sigma_i M_{i(j)}(\theta)).$$

Solving $\frac{\partial G}{\partial \lambda_j} = 0$ for $j = 1, \ldots, C$ and $\frac{\partial G}{\partial \theta_{1c}} = 0$ for $c = 1, \ldots, C$ gives

$$\Sigma_j \omega_{jc} n_{+j} = \Sigma_j \omega_{jc}\lambda_j \quad \text{for } c = 1, \ldots, C$$

which implies that $\lambda_j = n_{+j}$ for $j = 1, \ldots, C$.

The minimum discrepancy parameters are then found as the solution to

$$\Sigma_j \Sigma_i \psi_{ir}\omega_{jc}[n_{+j}M_{i(j)}(\theta^0(Q)) - n_{+j}\pi_{i(j)}] = 0 \quad \text{for } (r,c) \in Q. \quad \bullet$$

The system (3) is similar to its two-way multinomial counterpart, which prompts the following investigation. Consider "unconditioning" the operating models probabilities by multiplying each $\pi_{i(j)}$ by $\frac{n_{+i}}{n_{++}}$. (The transformed probabilities $\frac{n_{+i}}{n_{++}}\pi_{i(j)}$ then satisfy $\Sigma_j\Sigma_i\frac{n_{+i}}{n_{++}}\pi_{i(j)} = 1.$) These transformed probabilities might be modelled as

$$M_{ij}(\rho) = \exp\Big(\sum_{(r,c)\in Q}\psi_{ir}\omega_{jc}\rho_{rc}(Q)\Big) \quad \text{for all } i \text{ and } j$$

where the minimum discrepancy parameters are now found using the *multinomial* system

$$n_{++}\sum_j\sum_i\psi_{ir}\omega_{jc}\left[\exp\Big(\sum_{(r,c)\in Q}\psi_{ir}\omega_{jc}\rho^0_{rc}(Q)\Big) - \frac{n_{+j}}{n_{++}}\pi_{i(j)}\right] = 0 \quad \text{for } (r,c)\in Q. \tag{4}$$

The next theorem states that this is equivalent to solving (3), the product-multinomial system.

Theorem 3. Let $\theta^0_{rc}(Q)$ and $\rho^0_{rc}(Q)$ be defined by (3) and (4) respectively (for the same $Q \supseteq \{(1,c) : c = 1,\ldots,C\}$). Then

(a) $$\rho^0_{rc}(Q) = \begin{cases}\theta^0_{rc}(Q) + \sqrt{R}\sum_j \omega_{jc}\log\frac{n_{+i}}{n_{++}} & \text{for } r = 1\\ \theta^0_{rc}(Q) & \text{otherwise}\end{cases}$$

(b) $$\frac{n_{++}}{n_{+j}}M_{ij}(\rho^0) = M_{i(j)}(\theta^0) \quad \text{for all } i \text{ and } j$$

where

$$M_{ij}(\rho^0) = \sum_{(r,c)\in Q}\psi_{ir}\omega_{jc}\rho^0_{rc}(Q) \quad \text{for all } i \text{ and } j$$

$$M_{i(j)}(\theta^0) = \sum_{(r,c)\in Q}\psi_{ir}\omega_{jc}\theta^0_{rc}(Q) \quad \text{for all } i \text{ and } j.$$

Proof. Let $\rho^0_{rc}(Q)$ be defined by (4). Put

$$\theta^*_{rc}(Q) = \begin{cases} \rho^0_{rc}(Q) - \sqrt{R}\Sigma_k \omega_{kc} \log \frac{n_{+k}}{n_{++}} & \text{for } r = 1 \\ \rho^0_{rc}(Q) & \text{otherwise.} \end{cases}$$

Then, for all i and j

$$\begin{aligned} &\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\rho^0_{rc}(Q) \\ &= \sum_{c=1}^{C} \frac{1}{\sqrt{R}}\omega_{jc}(\theta^*_{1c}(Q) + \sqrt{R}\sum_k \omega_{kc} \log \frac{n_{+k}}{n_{++}}) + \sum_{(r,c)\in Q^*} \psi_{ir}\omega_{jc}\theta^*_{rc}(Q) \\ &= \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta^*_{rc}(Q) + \sum_k \log \frac{n_{+k}}{n_{++}} \left(\sum_{c=1}^{C} \omega_{jc}\omega_{kc}\right) \\ &= \sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta^*_{rc}(Q) + \log \frac{n_{+j}}{n_{++}} . \end{aligned}$$

Hence, for all i and j

$$\exp\left(\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\rho^0_{rc}(Q)\right) = \frac{n_{+j}}{n_{++}} \exp\left(\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta^*_{rc}(Q)\right)$$

and (4) can be written as

$$\sum_j \sum_i \psi_{ir}\omega_{jc}\left[n_{+j} \exp\left(\sum_{(r,c)\in Q} \psi_{ir}\omega_{jc}\theta^*_{rc}(Q)\right) - n_{+j}\pi_{i(j)}\right] = 0 \quad \text{for } (r,c) \in Q.$$

Since this is identical to (3) if one replaces $\theta^0_{rc}(Q)$ by $\theta^*_{rc}(Q)$ for each $(r,c) \in Q$, it follows that

$$\theta^0_{rc}(Q) = \theta^*_{rc}(Q) \quad \text{for all } (r,c) \in Q$$

which proves part (a). Part (b) then follows directly. •

This theorem allows one to use, for both multinomial and product-multinomial two-way classifications

(i) the same theoretical results regarding parameter and model interpretation, and

(ii) the same algorithm (with only minor modifications) for parameter estimation and model selection.

Furthermore the same applies to multiway tables. For the two-way fixed columns case considered, it was necessary that the product-multinomial models contain $\{\theta_{1c} \ : \ c = 1,\ldots,C\}$. In general one must simply ensure that for product-multinomial operating models the models fitted must contain the parameters corresponding to the marginals that are fixed; one can then proceed to treat the operating model as though it were multinomial, at least for the purpose of parameter estimation and model selection.

Estimating the expected discrepancy. For a given approximating family $\mathbf{M}(Q)$ the minimum discrepancy parameters are obtained by replacing $\pi_{i(j)}$ by n_{ij}/n_{+j} in the system of equations (3), which is then solved numerically. Let $\hat{\theta}(Q)$ denote the resulting estimates of the parameters. The discrepancy between the fitted model and the operating model is

$$\Sigma_j \left\{ -n_{+j} \Sigma_j \pi_{i(j)} \log M_{i(j)}(\hat{\theta}(Q);\ N) \right\} \tag{5}$$

where N is the matrix of cell counts. For each j, the term within the curly brackets is the overall discrepancy for a multinomial operating model, $\underline{\pi}_{(j)}$. Hence for each j, the expectation of this term (with respect to $\pi_{(j)}$) can be estimated by

$$-\frac{(n_{+j}-1)}{n_{+j}} \Sigma_i n_{ij} \log M_{i(j)}(\hat{\theta}(Q);\ N^{ij})$$

where, as before

(1) $n_{ij} \log M_{i(j)}(\hat{\theta}(Q);\ N^{ij})$ is defined to be zero if $n_{ij} = 0$; and

(2) N^{ij} is the matrix of cell counts where the (i,j)th entry has been decreased by 1.

Thus the expectation of (5) (with respect to π) is estimated by

$$-\Sigma_j \left\{ \frac{n_{+j}-1}{n_{+j}} \sum_i n_{ij} \log M_{i(j)}(\hat{\theta}(Q); N^{ij}) \right\}.$$

As in the multinomial case the cross-validated discrepancy can be written in orthogonal form, as

$$\sum_{(r,c)\in Q} \left(-\sum_i \sum_j \frac{n_{+j}-1}{n_{+j}} \psi_{ir}\omega_{jc} n_{ij}(\hat{\theta}_{rc}(Q); N^{ij}) \right)$$

where $n_{ij}(\hat{\theta}_{rc}(Q); N^{ij})$ is defined to be zero if $n_{ij} = 0$, and the $(\hat{\theta}_{rc}(Q); N^{ij})$ are the parameter estimates obtained when one observation has been removed from the (i,j)th cell.

The cross-validated discrepancies and the contributions of individual parameters can, for two-way multinomial and product-multinomial operating models, be obtained using practically the same algorithm.

In view of Theorem 3 and the similarities between the multinomial and product-multinomial cases, we will for the remainder of this chapter restrict attention to the multinomial case.

C. HIERARCHICAL MODELS FOR TWO-WAY TABLES

A special class of approximating families, called the hierarchical class is now considered. For models in this class, we will assume throughout that the joint model basis is constructed as the product of individual bases, one for each variable. The relationship between these models and the standard hierarchical loglinear models is discussed in D below.

The saturated model can be written as

$$\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{C}} \sum_{r=2}^{R} \psi_{ir}\theta_{r1} + \frac{1}{\sqrt{R}} \sum_{c=2}^{C} \omega_{jc}\theta_{1c} + \sum_{r=2}^{R}\sum_{c=2}^{C} \psi_{ir}\omega_{jc}\theta_{rc} \quad \text{for all } i,j$$

where the constant term, the two sets of main-effect parameters and the set of interaction parameters have been separated. Each of these sets are said to have a particular order. The constant term has the lowest order; the two sets of main-effect parameters (which both have the same order) have the next highest order while the set of interaction parameters have the next highest order.

The hierarchical class consists of approximating families which satisfy the following *hierarchy principle.* The principle given here holds for all dimensions. It has two parts:

(i) if one of the parameters in a given set (such as the set of row-effect parameters) is included, then all of the parameters within that set must also be included;

(ii) if a set of parameters of a particular order is included then all the related lower-order sets must also be included.

Examples

1. If a hierarchical model contains a single row-effect parameter, say θ_{21}, then it must contain all the row-effect parameters θ_{r1} for $r = 2,\ldots,R$; as well as the lower order term θ_{11}. It need not contain any other parameters.

2. If a hierarchical model contains an interaction parameter such as θ_{22}, then it must contain all the interaction parameters (θ_{rc} for $r = 2,\ldots,R$; $c = 2,\ldots,C$). The related lower order terms are the row and column effect parameters (θ_{r1} for $r = 2,\ldots,R$ and θ_{1c} for $c = 2,\ldots,C$), which must then also be included. The inclusion of the main-effect parameters then demands the inclusion of the lower order term θ_{11}.

The class of hierarchical approximating families for two-way tables contains a total of five families, each of which can be characterised by the sets of parameters which it contains. Let us define

$$\mathbf{R} = \{2,3,\ldots,R\}$$

$$\mathbf{C} = \{2,3,\ldots,C\}$$

$$\mathbf{R} \times \mathbf{C} = \{(r,c) \ : \ r \in \mathbf{R}, c \in \mathbf{C}\}.$$

Then the sets of parameters which each of the approximating families contain are

(1) θ_{11}

(2) θ_{11} and $\{\theta_{r1}\}_{r\in\mathbf{R}}$

(3) θ_{11} and $\{\theta_{1c}\}_{c\in\mathbf{C}}$

(4) θ_{11}, $\{\theta_{r1}\}_{r\in\mathbf{R}}$ and $\{\theta_{1c}\}_{c\in\mathbf{C}}$

(5) θ_{11}, $\{\theta_{r1}\}_{r\in\mathbf{R}}$, $\{\theta_{1c}\}_{c\in\mathbf{C}}$ and $\{\theta_{rc}\}_{r\in\mathbf{R};\, c\in\mathbf{C}}$.

The list below gives, for each of the families,

(a) a typical model from the family

(b) special properties of the modelled probabilities

(c) the corresponding interpretation.

1.(a) $\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11}$ for all i,j

(b) $M_{ij}(\theta) = \frac{1}{RC}$ for all i,j

(c) Both variables redundant.

2.(a) $\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{C}}\sum_{r=2}^{R}\psi_{ir}\theta_{r1}$ for all i,j

(b) $M_{ij}(\theta) = M_{i+}(\theta)/C$ for all i,j

(c) The column variable redundant.

3.(a) $\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{R}}\sum_{c=2}^{C}\omega_{jc}\theta_{1c}$ for all i,j

(b) $M_{ij}(\theta) = M_{+j}(\theta)/R$ for all i,j

(c) The row variable redundant.

4.(a) $\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{C}}\sum_{r=2}^{R}\psi_{ir}\theta_{r1} + \frac{1}{\sqrt{R}}\sum_{c=2}^{C}\omega_{jc}\theta_{1c}$ for all i,j

(b) $M_{ij}(\theta) = M_{i+}(\theta)M_{+j}(\theta)$ for all i,j

(c) The row and column variables are independent.

5.(a) $\log M_{ij}(\theta) = \frac{1}{\sqrt{RC}}\theta_{11} + \frac{1}{\sqrt{C}}\sum_{r=2}^{R}\psi_{ir}\theta_{r1} + \frac{1}{\sqrt{R}}\sum_{c=2}^{C}\omega_{jc}\theta_{1c}$
$+ \sum_{r=2}^{R}\sum_{c=2}^{C}\psi_{ir}\omega_{jc}\theta_{rc}$ for all i,j

(b) $M_{ij}(\theta^0) = \pi_{ij}$ for all i,j

(c) The saturated model.

For each of the hierarchical approximating families it is possible to obtain explicit expressions for the minimum discrepancy parameters in terms of the operating model probabilities or their marginals. These can be obtained via a two-step process.

Firstly the optimal model from any hierarchical model can be written as

$$\log M_{ij}(\theta^0(Q)) = \frac{1}{\sqrt{RC}}\theta^0_{11}(Q) + \chi_{\mathbf{R}\times 1}(Q)\frac{1}{\sqrt{C}}\sum_{r=2}^{R}\psi_{ir}\theta^0_{r1}(Q)$$
$$+ \chi_{1\times \mathbf{C}}(Q)\frac{1}{\sqrt{R}}\sum_{c=2}^{C}\omega_{jc}\theta^0_{1c}(Q) + \chi_{\mathbf{R}\times\mathbf{C}}(Q)\sum_{r=2}^{R}\sum_{c=2}^{C}\psi_{ir}\omega_{jc}\theta^0_{rc}(Q) \text{ for all } i,j \quad (6)$$

where

$$\mathbf{R}\times 1 = \{(r,1) : r\in\mathbf{R}\}, \quad 1\times\mathbf{C} = \{(1,c) : c\in\mathbf{C}\}$$

and

$$\chi_{A\times B}(Q) = \begin{cases} 1 & \text{if } A\times B\subseteq Q \\ 0 & \text{otherwise.}\end{cases}$$

Multiplying (6) by $\psi_{ir}\omega_{jc}$ throughout and then summing over all i and j gives

$$\theta^0_{rc}(Q) = \Sigma_i\Sigma_j\psi_{ir}\omega_{jc}\log M_{ij}(\theta^0(Q)) \quad \text{for each } r,c\in Q. \quad (7)$$

The second step involves substituting for $M_{ij}(\theta^0(Q))$ in (7). For each different approximating family a different substitution is appropriate. The list below gives, for each of the five hierarchical families in turn, the corresponding expression for $M_{ij}(\theta^0(Q))$.

(1) Both variables redundant

$$M_{ij}(\theta^0(Q)) = \frac{1}{RC} \quad \text{for all } i,j$$

(2) Column variable redundant

$$M_{ij}(\theta^0(Q)) = \pi_{i+}/C \quad \text{for all } i,j$$

(3) Row variable redundant

$$M_{ij}(\theta^0(Q)) = \pi_{+j}/R \quad \text{for all } i,j$$

(4) Row and column variables independent

$$M_{ij}(\theta^0(Q)) = \pi_{i+}\pi_{+j} \quad \text{for all } i,j$$

(5) The saturated family

$$M_{ij}(\theta^0(Q)) = \pi_{ij} \quad \text{for all } i,j.$$

As an example of how these equalities were derived consider the independence family. For models in this family we have that

$$M_{ij}(\theta^0) = M_{i+}(\theta^0)M_{+j}(\theta^0) \quad \text{for all } i,j.$$

Furthermore, from Theorem 1 we have, since the model contains $\{\theta^0_{r1}\}_{r=1,\ldots,R}$ and $\{\theta_{1c}\}_{c=1,\ldots,C}$, that

$$M_{i+}(\theta^0) = \pi_{i+} \quad \text{for all } i$$

$$M_{+j}(\theta^0) = \pi_{+j} \quad \text{for all } j.$$

Combining these two sets of results gives the equality

$$M_{ij}(\theta^0) = \pi_{i+}\pi_{+j} \quad \text{for all } i,j.$$

When the expressions for the $M_{ij}(\theta^0(Q))$ given in the above list are substituted into (7) the expressions for the optimal parameters can be simplified in some cases. The list below gives the simplified expressions for the optimal parameters in each of the hierarchical families.

(1) Both variables redundant

$$\theta^0_{11}(Q) = \sqrt{RC}\,\log(\frac{1}{RC})$$

(2) Column variable redundant

$$\theta^0_{11}(Q) = \sqrt{\frac{C}{R}}\,\Sigma_i\,\log(\pi_{i+}/C)$$
$$\theta^0_{r1}(Q) = \sqrt{C}\,\Sigma_i\psi_{ir}\,\log(\pi_{i+}/C) \quad \text{for } r \in \mathbf{R}$$

(3) Row variable redundant

$$\theta^0_{11}(Q) = \sqrt{\frac{R}{C}}\,\Sigma_i\,\log(\pi_{+j}/R)$$
$$\theta^0_{1c}(Q) = \sqrt{R}\,\Sigma_j\omega_{jc}\,\log(\pi_{+j}/R) \quad \text{for } c \in \mathbf{C}$$

(4) Row and column variables independent

$$\theta^0_{11}(Q) = \frac{1}{\sqrt{RC}}\,\Sigma_i\Sigma_j\,\log(\pi_{i+}\pi_{+j})$$
$$\theta^0_{r1}(Q) = \sqrt{C}\,\Sigma_i\psi_{ir}\,\log\,\pi_{i+} \quad \text{for } r \in \mathbf{R}$$
$$\theta^0_{1c}(Q) = \sqrt{R}\,\Sigma_j\omega_{jc}\,\log\,\pi_{+j} \quad \text{for } c \in \mathbf{C}$$

(5) The saturated family

$$\theta^0_{11}(Q) = \frac{1}{\sqrt{RC}}\,\Sigma_i\Sigma_j\,\log\,\pi_{ij}$$
$$\theta^0_{r1}(Q) = \frac{1}{\sqrt{C}}\,\Sigma_i\psi_{ir}(\Sigma_j\,\log\,\pi_{ij}) \quad \text{for } r \in \mathbf{R}$$
$$\theta^0_{1c}(Q) = \frac{1}{\sqrt{R}}\,\Sigma_j\omega_{jc}(\Sigma_i\,\log\,\pi_{ij}) \quad \text{for } c \in \mathbf{C}$$
$$\theta^0_{rc}(Q) = \Sigma_i\Sigma_j\psi_{ir}\omega_{jc}\,\log\,\pi_{ij} \quad \text{for } (r,c) \in \mathbf{R} \times \mathbf{C}.$$

These parameters are estimated, as always, by replacing each operating model quantity with its sample analog. Thus for hierarchical models the parameters can be estimated directly.

D. STANDARD HIERARCHICAL MODELS FOR TWO-WAY TABLES

From the foregoing it is clear that hierarchical models have very clear and simple interpretations. It is thus not surprising that there is a large literature on the subject (see, for example, Goodman (1970), Plackett (1974), Bishop *et al* (1975), Fienberg (1977) to mention only a few).

These authors use different parameterisations than those used in basis models. For a two-way table consider the parameterisation

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad \text{for all } i,j \tag{8}$$

where the m_{ij} are the expected cell proportions under the model. As a model (8) is over-parameterised involving $(1+R+C+RC)$ parameters for RC cells. The parameterisation is similar to that of an analysis-of-variance (ANOVA) model for a two-way cross-classification. This suggests the ANOVA-type constraints

$$\begin{aligned}
\Sigma_i u_{1(i)} &= 0 \\
\Sigma_j u_{2(j)} &= 0 \\
\Sigma_i u_{12(ij)} &= 0 \quad \text{for} \quad j = 1,\ldots,C \\
\Sigma_j u_{12(ij)} &= 0 \quad \text{for} \quad i = 1,\ldots,R.
\end{aligned}$$

The parameterisation (8) together with these $(2+R+C)$ restrictions then defines the standard saturated model.

As with basis models the parameters are formed into sets. It is usual to define

$$\begin{aligned}
\{u_{1(i)}\}_{i=1,\ldots,R} &= \mathbf{u}_1 \\
\{u_{2(j)}\}_{j=1,\ldots,C} &= \mathbf{u}_2 \\
\{u_{12(ij)}\}_{i=1,\ldots,R;\, j=1,\ldots,C} &= \mathbf{u}_{12}
\end{aligned}$$

where $\mathbf{u}_1$ contains the row-effect parameters, $\mathbf{u}_2$ the column-effect parameters and $\mathbf{u}_{12}$ the (first-order) row $\star$ column interaction parameters. The sets of parameters are then ordered in the obvious way and the standard hierarchical models are constructed in accordance with the hierarchy principle given earlier. The only

difference being that now if a parameter set is excluded then the restrictions involving these parameters must also be dropped. As with basis models there are five hierarchical models (for two-way tables). The list below gives for each of the five

(a) the model

(b) any special properties

(c) the interpretation.

1.(a) $\log m_{ij} = u$ for all i,j (with $\Sigma_i m_{ij} = 1$)

(b) $m_{ij} = \frac{1}{RC}$ for all i,j

(c) Both variables redundant.

2.(a) $\log m_{ij} = u + u_{1(i)}$ for all i,j with the constraint $\Sigma_i u_{1(i)} = 0$

(b) $m_{ij} = m_{i+}/C$ for all i,j

(c) Column variable redundant.

3.(a) $\log m_{ij} = u + u_{2(j)}$ for all i,j with the constraint $\Sigma_j u_{2(j)} = 0$

(b) $m_{ij} = m_{+j}/R$ for all i,j

(c) Row variable redundant.

4.(a) $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$ for all i,j with the constraints

$$\Sigma_i u_{1(i)} = \Sigma_j u_{2(j)} = 0$$

(b) $m_{ij} = m_{i+}m_{+j}$ for all i,j

(c) Row and column variables independent.

5.(a) $\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ for all i,j with the constraints

$$\Sigma_i u_{1(i)} = \Sigma_j u_{2(j)} = 0$$

$$\Sigma_i u_{12(ij)} = \Sigma_j u_{12(ij)} = 0$$

(b) -

(c) The saturated model.

Note that the interpretation of each of these five models is identical to that of the corresponding basis model. There is obviously a very close link between these models and the corresponding class of basis models. In fact, (provided attention is restricted to basis models whose joint model is obtained as the product of individual bases) it can be shown that, for any given basis, there is a one-to-one correspondence between standard hierarchical model and each basis hierarchical model. Furthermore it is possible to give the relationship between the parameters in each of the two types of parameterisations. For the case of two-way tables it can be shown that (up to scaling factors):

$$\theta_{11} = u$$

$$\theta_{r1} = \Sigma_i \psi_{ir} u_{1(i)} \quad \text{for } r = 2, \ldots, R$$

$$\theta_{1c} = \Sigma_j \omega_{jc} u_{2(j)} \quad \text{for } c = 2, \ldots, C$$

$$\theta_{rc} = \Sigma_i \Sigma_j \psi_{ir} \omega_{jc} u_{12(ij)} \quad \text{for } r = 2, \ldots, R;\ c = 2, \ldots, C.$$

The inverse relationships are simply:

$$u = \theta_{11}$$

$$u_{1(i)} = \Sigma_i \psi_{ir} \theta_{r1} \quad \text{for } i = 1, \ldots, R$$

$$u_{2(j)} = \Sigma_j \omega_{jc} \theta_{1c} \quad \text{for } j = 1, \ldots, C$$

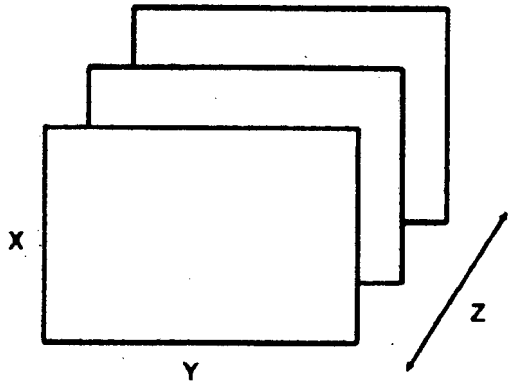
$$u_{12(ij)} = \Sigma_i \Sigma_j \psi_{ir} \omega_{jc} \theta_{rc} \quad \text{for } i = 1, \ldots, R;\ j = 1, \ldots, C.$$

5.3 MULTIWAY CROSS-CLASSIFICATIONS

We now consider tables which involve more than two variables. For some purposes, such as the estimation of the parameters and the estimation of the expected discrepancy, it is convenient to simply regard these tables as either one- or two-way tables for which the details were given in the previous two sections. However, the introduction of additional variables brings with it the possibility of describing various forms of independence between the variables or groups thereof, and it is on this that we will concentrate in this section.

THREE-WAY CROSS-CLASSIFICATIONS

Consider a cross-classification involving three variables X, Y and Z ; which can be arranged as a stack of two-way tables.



We will refer to X, Y and Z as the row-,column- and layer- variable respectively which will be taken to have R, C and L categories respectively. The (i, j, k)th cell will refer to the cell in the ith row of the jth column and the kth layer.

If $[\psi_{ir}]_{i,r}, [\omega_{jc}]_{j,c}$ and $[\phi_{k\ell}]_{k,\ell}$ are model bases for X, Y and Z respectively then a typical model is of the form

$$M(\theta) = [M_{ijk}(\theta)]_{i,j,k}$$

where

$$\log M_{ijk}(\theta) = \sum_{(r,c,\ell)\in Q} \psi_{ir}\omega_{jc}\phi_{k\ell}\theta_{rc\ell}$$

and the model must contain the set of parameters corresponding to the marginal totals which are considered fixed. For example, if the layer totals are considered fixed then the model must contain $\{\theta_{11\ell}\}_{\ell=1,\ldots,L}$.

The parameters have their usual interpretations. For example,

(1) $\{\theta_{r11}\}_{r\in\mathbf{R}}, \{\theta_{1c1}\}_{c\in\mathbf{C}}$ and $\{\theta_{11\ell}\}_{\ell\in\mathbf{L}}$ are the three sets of main-effect parameters for rows, columns and layers respectively,

(2) $\{\theta_{rc1}\}_{(r,c)\in\mathbf{R}\times\mathbf{C}}, \{\theta_{r1\ell}\}_{(r,\ell)\in\mathbf{R}\times\mathbf{L}}$ and $\{\theta_{1c\ell}\}_{(c,\ell)\in\mathbf{C}\times\mathbf{L}}$ are the three sets of two-factor (first order) interaction parameters,

(3) $\{\theta_{rc\ell}\}_{(r,c,\ell)\in\mathbf{R}\times\mathbf{C}\times\mathbf{L}}$ are the three-factor (second order) interaction parameters.

Hierarchical models for three-way cross-classifications

In two-way tables the model of independence between the two variables was of particular interest. The introduction of a third variable introduces the possibility of further forms of independence between the variables or groups thereof. These can be modelled with the hierarchical models for three-way cross-classifications.

The saturated model can be written as

$$\begin{aligned}
&\log M_{ijk}(\theta) \\
&= \frac{1}{\sqrt{RCL}}\theta_{111} + \left(\frac{1}{\sqrt{CL}}\sum_{r=2}^{R}\psi_{ir}\theta_{r11} + \frac{1}{\sqrt{RL}}\sum_{c=2}^{C}\omega_{jc}\theta_{1c1} + \frac{1}{\sqrt{RC}}\sum_{\ell=2}^{L}\phi_{k\ell}\theta_{11\ell}\right) \\
&\quad + \left(\frac{1}{\sqrt{L}}\sum_{r=2}^{R}\sum_{c=2}^{C}\psi_{ir}\omega_{jc}\theta_{rc1} + \frac{1}{\sqrt{C}}\sum_{r=2}^{R}\sum_{\ell=2}^{L}\psi_{ir}\phi_{k\ell}\theta_{r1\ell} + \frac{1}{\sqrt{R}}\sum_{c=2}^{C}\sum_{\ell=2}^{L}\omega_{jc}\phi_{k\ell}\theta_{1c\ell}\right) \\
&\quad + \left(\sum_{r=2}^{R}\sum_{c=2}^{C}\sum_{\ell=2}^{L}\psi_{ir}\omega_{jc}\phi_{k\ell}\theta_{rc\ell}\right).
\end{aligned}$$

The hierarchical class of approximating families consists of those families which satisfy the hierarchy principle given in the previous section. A list of all the possible families is given below where each is represented by the parameters which it contains. In the list $\{\theta_{r11}\}_{r\in\mathbf{R}}$ is abbreviated to $\{\theta_{r11}\}$, $\{\theta_{rc1}\}_{(r,c)\in\mathbf{R}\times\mathbf{C}}$ to $\{\theta_{rc1}\}$, etc. In addition, ME is used as an abbreviation for the set of all main-effect

section. For the remaining cases we will only look at the (a) subcase where subcases are involved.

A list giving the special interpretation properties for each of the approximating families is given below. (The proofs of the quoted properties are straightforward and are therefore omitted.)

(5) **Complete independence.** X, Y and Z are completely independent

$$M_{ijk}(\theta) = M_{i++}(\theta)M_{+j+}(\theta)M_{++k}(\theta) \quad \text{for all } i, j \text{ and } k$$

(6a) **Joint independence.** X and Y are (dependent on each other but) jointly independent of Z

$$M_{ijk}(\theta) = M_{ij+}(\theta)M_{++k}(\theta) \quad \text{for all } i, j \text{ and } k$$

(7a) **Conditional independence.** Conditional on each given value of Z, X and Y are independent

$$M_{ijk}(\theta) = M_{i+k}(\theta)M_{+jk}(\theta)/M_{++k}(\theta) \quad \text{for all } i, j \text{ and } k$$

(8) **No second-order interaction.** There is pairwise first-order interaction among the three variables but no second-order interaction involving all three variables simultaneously. It is not possible to express $M_{ijk}(\theta)$ simply in terms of the marginals $\{M_{ij+}(\theta)\}$, $\{M_{i+k}(\theta)\}$ and $\{M_{+jk}(\theta)\}$.

(9) **The saturated model.** $M_{ijk}(\theta^0(Q)) = \pi_{ijk}$ for all i, j and k.

As in the two-dimensional case it is possible to obtain explicit expressions for the minimum discrepancy parameters (firstly) in terms of the $M_{ijk}(\theta^0(Q))$, namely

$$\theta^0_{rc\ell}(Q) = \Sigma_i\Sigma_j\Sigma_k\psi_{ir}\omega_{jc}\phi_{k\ell} \log M_{ijk}(\theta^0(Q)) \quad \text{for all } (r, c, \ell) \in Q.$$

For all of the models, with the exception of the one with no second-order interaction, it is possible to express $M_{ijk}(\theta^0(Q))$ in terms of the marginals or

individual cell probabilities in the operating model. The minimum discrepancy parameters can be written as:

(5) X, Y **and** Z **completely independent**

$$\theta^0_{111}(Q) = \frac{1}{\sqrt{RCL}} \Sigma_i\Sigma_j\Sigma_k \log(\pi_{i++}\pi_{+j+}\pi_{++k})$$
$$\theta^0_{r11}(Q) = \sqrt{CL}\, \Sigma_i\psi_{ir} \log \pi_{i++} \quad \text{for } r \in \mathbf{R}$$
$$\theta^0_{1c1}(Q) = \sqrt{RL}\, \Sigma_j\omega_{jc} \log \pi_{+j+} \quad \text{for } c \in \mathbf{C}$$
$$\theta^0_{11\ell}(Q) = \sqrt{RC}\, \Sigma_k\phi_{k\ell} \log \pi_{++k} \quad \text{for } \ell \in \mathbf{L}$$

(6a) X **and** Y **jointly independent of** Z

$$\theta^0_{111}(Q) = \frac{1}{\sqrt{RCL}} \Sigma_i\Sigma_j\Sigma_k \log(\pi_{ij+}\pi_{++k})$$
$$\theta^0_{rc1}(Q) = \sqrt{L}\, \Sigma_i\Sigma_j\psi_{ir}\omega_{jc} \log(\pi_{ij+}) \quad \text{for } r = 1,\ldots,R; c = 1,\ldots C; (r,c) \neq (1,1)$$
$$\theta^0_{11\ell}(Q) = \sqrt{RC}\, \Sigma_k\phi_{k\ell} \log(\pi_{++k}) \quad \text{for } \ell \in \mathbf{L}$$

(7a) X **and** Y **conditionally independent of** Z

$$\theta^0_{11\ell}(Q) = \frac{1}{\sqrt{RC}} \Sigma_k\phi_{k\ell}\{\Sigma_i\Sigma_j \log(\pi_{i+k}/\pi_{++k})\} \quad \text{for } \ell = 1,\ldots,L$$
$$\theta^0_{r1\ell}(Q) = \sqrt{C}\, \Sigma_i\Sigma_k\psi_{ir}\phi_{k\ell} \log(\pi_{i+k}) \quad \text{for } r = 2,\ldots,R;\ \ell = 1,\ldots,L$$
$$\theta^0_{1c\ell}(Q) = \sqrt{R}\, \Sigma_j\Sigma_k\omega_{jc}\phi_{k\ell} \log(\pi_{+jk}) \quad \text{for } c = 2,\ldots,C;\ \ell = 1,\ldots,L$$

(9) **The saturated model**

$$\theta^0_{rc\ell}(Q) = \Sigma_i\Sigma_j\Sigma_k\psi_{ir}\omega_{jc}\phi_{k\ell} \log \pi_{ijk} \quad \text{for } r = 1,\ldots,R;\ c = 1,\ldots C;\ \ell = 1,\ldots,L.$$

MORE THAN THREE VARIABLES

Loglinear basis models, like all basis models, are easily constructed for tables involving any number of variables. Essentially one simply adds a model basis for each additional variable. Furthermore the hierarchy principle applies to loglinear basis models having any number of variables, and can generate classes of approximating families containing models which can be used to model various forms of

independence between groups of variables in a particular cross-classification. However it is often difficult to think in terms of more than three variables and for tables involving more than three variables it is often convenient to regard a group of variables as a single variable (c.f. the analysis of the data sets in Section 4.2 and also in Section 5.5). •

5.4 QUASI-HIERARCHICAL MODELS

Recall that members of the hierarchical class of models must satisfy two conditions (see Section 5.2):

(i) if one of the parameters in a given set (such as the set of row effect parameters) is included then all the parameters within that set must also be included;

(ii) if a set of parameters of a particular order is included then all the related lower-order sets must also be included.

The first of these rules originates from the standard parameterisations of log-linear models. The rule makes good sense in this context because these parameters occur in sets which are such that the parameters within a set have to be modelled and interpreted jointly.

Consider for example the set of parameters $\{u_{1(i)} : i = 1,\ldots,R\}$ in any of the models (2), (4) or (5) in D of Section 5.2. As the indexing suggests, each $u_{1(i)}$ refers to an individual row mean. The $u_{1(i)}$ are linked via the constraint $\Sigma_i u_{i(1)} = 0$. In fact

$$u_{1(i)} = \frac{1}{C}\sum_j \log m_{ij} - \Big(\frac{1}{RC}\sum_i\sum_j \log m_{ij}\Big) \quad \text{for } i = 1,\ldots,R.$$

Thus each $u_{1(i)}$ measures the deviation of the ith (log) row mean from the grand (log) mean, and the value of any individual $u_{1(i)}$ affects the value of the grand mean and hence the values of the other $u_{1(i)}$.

Having to include or exclude whole groups of parameters means, in the context of two-way tables, that there are essentially only two hierarchical models of interest (the other hierarchical models involve at least one redundant variable). These two hierarchical models are (i) the saturated model and (ii) the model of independence between the two variables. It would clearly be useful to have models which lie "between" these two extremes; models which admit that the variables are dependent but which *model* the nature of the dependence between the two variables using fewer parameters than the saturated model. One could model that the two variables were related, for example, in a linear fashion.

It is in this context that basis models can be particularly useful. In basis models each parameter can be interpreted and considered for exclusion *separately.* This is a consequence of the property of basis models that each parameter has its own coefficient matrix which determines the nature of the parameter's contribution to the fitted probabilities and hence the parameter's interpretation. This feature of basis models cannot be exploited if one applies rule (i), which obliges one to deal with sets of parameters rather than with individual parameters.

Consequently when working with basis models we will consider models which satisfy (ii), but not necessarily (i) above. We will use the term *quasi-hierarchical* to describe this class of approximating families.

Examples. 1. A quasi-hierarchical basis model for a two-way classification which contains the interaction parameter θ_{23} , must contain the corresponding row- and column- main-effect parameters θ_{21} and θ_{13} , which in turn implies the inclusion of the constant term θ_{11} . (An hierarchical model on the other hand which contained θ_{23} would have to be the saturated model.)

2. A quasi-hierarchical basis model for a three-way table which contains the three-factor interaction parameter θ_{222} , must also contain the corresponding two-factor interaction parameters $\theta_{122}, \theta_{212}$ and θ_{221} ; together with the corresponding main-effect parameters $\theta_{112}, \theta_{121}$ and θ_{211} ; as well as the constant parameter θ_{111} .

The quasi-hierarchical class provides, besides all the hierarchical models, models which are between the hierarchical models, modelling with a few parameters the nature of the dependence, if any, between the variables in a cross-classification, rather than simply modelling either complete dependence or independence; at the same time remaining easy to interpret.

Since the class of quasi-hierarchical models contains that of hierarchical models it follows that in the former case, in performing model selection, we need to examine a larger number of models. This increases the computational load. However (in this context) this disadvantage is easily outweighed by the additional flexibility gained.

In the examples which follow, the selected models (which are selected from the quasi-hierarchical class) are mostly not hierarchical. They generally contain fewer parameters than the best hierarchical model for the same data set. •

5.5 EXAMPLES

We now fit loglinear models to the six data sets introduced in Section 3.2 and to which linear models were fitted in Section 4.2. In each case we use the same bases as before. In those cases where two possible bases for a single variable were given previously we will now only consider the more successful of the two.

Two general features which emerge when the loglinear models are fitted, are:

(a) the linear and the loglinear basis model fitted to each of the data sets are in general similar; both in terms of the actual fitted cell probabilities and in terms of the parameters which the selected fitted models contain,

(b) except for tables which, like the lizard data set, involve variables with a few categories each, the models selected although they are chosen to be quasi-hierarchical are not hierarchical.

THE TREATMENT DATA. Recall the treatment data where two treatments in one of the sequences AB, BA, AA, and BB were administered to patients who were asked to indicate their preference ("prefers first", "prefers second", "no preference").

The model bases used previously for the preference and sequence variables are, respectively,

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -2 & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Let $M_{i(j)}(\theta)$ denote the modelled probability a randomly selected patient who receives the jth treatment combination will indicate the ith preference category $(i = 1, \ldots, 3;\ j = 1, \ldots, 4)$. The models considered are of the form:

$$\log M_{i(j)(\theta)} = \frac{1}{\sqrt{3}} \sum_{c=1}^{4} \omega_{jc}\theta_{1c} + \sum_{(r,c)\in Q^*} \psi_{ir}\omega_{jc}\theta_{rc}$$

where

(a) $Q^* \subseteq \{(r,c) : r = 1,2,3;\ c = 2,3,4\}$

(b) $M_{+(j)}(\theta) = 1$ for $j = 1,\ldots,4$

(c) $[\psi_{ir}]_{i=1,2,3;\ r=1,2,3}$ and $[\omega_{jc}]_{j=1,\ldots,4;\ c=1,\ldots,4}$ are the normalised model bases.

The θ_{1c} have to be included in the model because they are the parameters which correspond to the conditioning variable's marginals (i.e. the sequence variable's marginals).

The modelling procedure is begun by fitting the saturated model. The contributions ($\times 10^3$) to the cross-validated discrepancy are shown in the following *contribution table.*

	c = 1	c = 2	c = 3	c = 4	
r=1	221.87	0.62	-0.07	-0.01	
r=2	-48.21	-2.97	1.09	1.02	
r=3	-3.96	0.97	-5.94	1.05	Total: 165.44

The contributions in the first row correspond to those parameters which have to be included in the model. Of the remaining parameters there are four candidates for exclusion, namely the four parameters whose contributions are positive. The exclusion conveniently leads to a quasi-hierarchical model. After re-estimating the remaining parameters the contribution table for this reduced model is:

	c=1	c=2	c=3	c=4	
r=1	221.28	0.50	-0.01	0.00	
r=2	-47.95	-2.84	***	***	
r=3	-3.82	***	-5.93	***	Total: 161.16

All the remaining optional parameters assist in decreasing the cross-validated discrepancy, and so, according to this criterion, no further exclusion of parameters is appropriate. That the corresponding model indeed leads to the smallest criterion was confirmed by examining all possible models.

The list below gives a short description of each of the parameters in the selected model:

- θ_{21} – row-effect parameter, involving the "preference-or-not" contrast
- θ_{31} – row-effect parameter, involving the "prefer-first-or-not" contrast
- θ_{22} – interaction parameter; (preference-or-not) $\star$ (two-treatments-or-not)
- θ_{33} – interaction parameter; (prefer-first-or-not) $\star$ (AB-or-BA).

The estimated parameter values ($\times 1000$)

	c=1	c=2	c=3	c=4
r=1	-4482.4	155.0	-134.1	-0.02
r=2	-1830.3	416.1	$\star\star\star$	$\star\star\star$
r=3	774.6	$\star\star\star$	889.6	$\star\star\star$

As with linear models the sign of each estimated parameter considered in conjunction with its interaction matrix indicates "the direction of the trend". For example, that $\theta_{21} = -1,8303$ is negative and its interaction matrix is

$$\frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \otimes \frac{1}{2}\underline{1}'$$

suggests that, averaging over the four treatment sequences, most patients indicated "no-preference"; which is not surprising as two of the treatment sequences involve giving patients the same treatment twice. Similarly $\hat{\theta}_{33} = 0,8896$ is positive and its interaction matrix is

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \otimes \frac{1}{\sqrt{2}}(1,-1,0,0) = \begin{pmatrix} \frac{1}{2} & 1 & -1 & 0 & 0 \\ -1 & & 1 & 0 & 0 \\ 0 & & 0 & 0 & 0 \end{pmatrix}$$

suggesting that more patients prefer the first treatment when given AB, while more prefer the second when given BA; thus indicating a general preference for A among those who received two different treatments.

It should be noted that the selected linear and loglinear model both contain the same four parameters and therefore the same interaction matrices. Thus the same features in the data set emerge in both cases. •

THE LIZARD DATA SET. The lizard data comprised counts of the number of lizards at different (perch height, perch diameter) values for two species. Of interest are the (height, diameter) probabilities conditional on species. All the variables are binary and the Hadamard basis is used for each. There are eight possible parameters. These can be written in the format:

θ_{111}	θ_{121}
θ_{211}	θ_{221}

θ_{112}	θ_{122}
θ_{212}	θ_{222}

Writing H, D and S to represent *height, diameter* and *species* respectively, the following table indicates the interpretation of each of the above parameters.

+	D
H	H ⋆ D

S	D ⋆ S
H ⋆ S	H ⋆ D ⋆ S

Since the probabilities are conditional on species, all models must contain $\theta_{11\ell}$ for $\ell = 1, 2$.

For the saturated model the contributions $(\times 10^3)$ of each parameter to the cross-validated discrepancy are :

619.05	-27.04
-41.31	1.85

-8.94	-8.28
-5.56	-0.08

There are two candidates for exclusion namely θ_{221} and θ_{222} (the latter because its contribution is very small albeit negative). There are only two quasi-hierarchical models which can be obtained by excluding either one or both of these variables; namely

(1) the model with only θ_{222} excluded – the "no second-order interaction" model, and

(2) the model with both θ_{222} and θ_{221} excluded – which means that of the two-factor interaction parameters the model contains H $\star$ S and D $\star$ S but not H $\star$ D. This is the model of conditional independence of height and diameter for each species.

Their respective contribution tables are given by:

The contribution table (no second-order interaction)

620.18	-28.08	-9.37	-8.76	
-42.15	1.97	-5.88	$\star\star\star$	Total: 528.34

The contribution table (height and diameter independent for each species)

615.94	-25.70	-8.76	-8.17	
-39.61	$\star\star\star$	-5.39	$\star\star\star$	Total: 528.30

The second of these models has a slightly lower cross-validated discrepancy and also the simpler interpretation. •

THE CAMP DATA. The data are counts of the responses of U.S. Army Recruits to a questionnaire. There are three explanatory variables; namely the location of the current training camp (L), the geographic region of origin of the recruit (O) and the race of the recruit (R). The joint model basis $[\omega_{jc}]_{j=1,\ldots,8;\ c=1,\ldots,8}$ used for the three binary explanatory variables is obtained by re-ordering the columns of H_8 so that the columns can be identified as

$$+,\ L,\ O,\ R,\ L\star O,\ L\star R,\ O\star R,\ L\star O\star R$$

respectively. The five response categories and the (non-normalised) model basis $[\psi_{ir}]_{i=1,\ldots,5;\ r=1,\ldots,5}$ are:

prefer to stay		1	1	0	1	0
prefer	North	1	1	0	-1/2	1
to	South	1	1	0	-1/2	-1
move	undecided	1	-3/2	1	0	0
undecided		1	-3/2	-1	0	0

In this application we wish to model the conditional probabilities of the response categories for each of the given (composite) explanatory variable categories. The models considered have the form

$$M_{ij}(\theta) = \exp\left(\frac{1}{\sqrt{5}}\sum_{c=1}^{8}\omega_{jc}\theta_{rc} + \sum_{(r,c)\in Q^*}\psi_{ir}\omega_{jc}\theta_{1c}\right) \quad \text{for } i = 1,\ldots,5;\ j = 1,\ldots,8$$

with $M_{+(j)}(\theta) = 1$ for $j = 1,\ldots 8$, and $Q^* \subseteq \{(r,c) : r = 2,\ldots,5 : c = 1,\ldots,8\}$. (Although the models are written in terms of only two subscripts it should be borne in mind, particularly in determining whether a model is quasi-hierarchical, that the column variable subscripts, j and c, in fact each represent three variables.)

The contribution table (saturated model)

	+	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_1$	18785.6	165.0	4.7	5.8	-12.4	-3.9	-270.2	-10.3
$\underline{\psi}_2$	-1324.5	-9.7	1.0	2.1	-32.0	16.0	29.1	1.0
$\underline{\psi}_3$	-16.6	-13.0	-1.1	0.7	-1.7	-0.1	-10.3	1.1
$\underline{\psi}_4$	-1.8	-8.3	-172.8	2.0	-450.5	-108.7	-40.4	-20.8
$\underline{\psi}_5$	-1301.4	-5.3	-1956.2	40.2	-38.6	2.5	-62.7	15.7

Total: 14426.87

Remarks. 1. A remarkable feature about this table is that all the race ⋆ preference parameters have positive contributions. This indicating that race, on its own, is irrelevant to the proportions in each of the preference categories. However there are some higher order parameters which involve race, in particular the origin ⋆ race ⋆ preference parameters, which do make negative contributions. For the purposes of preliminary selection it must be remembered that quasi-hierarchical models cannot contain an $O \star R \star \underline{\psi}_r$ parameter without including both the $O \star \underline{\psi}_r$ and the $R \star \underline{\psi}_r$ parameters.

2. Apart from the contributions in the first column (which are associated with parameters which average over the three explanatory variables), the largest negative contribution is due to the O ⋆ (move-North-or-South contrast) parameter. This once

again indicates that the major trend is for recruits to want to move to a camp which is closer to home.

As part of the model selection process a number of different models were investigated. It became apparent that models which excluded only a few parameters generally lead to low discrepancies. In part this can be explained by the large sample size which allows a greater number of model parameters to be estimated accurately. In addition there appears to be strong interaction between the variables.

The contribution tables of some of the more successful models are given below.

The contribution tables

Model I

	+	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_1$	18770.2	163.2	4.5	3.1	-19.7	11.4	-268.4	-10.6
$\underline{\psi}_2$	-1330.0	-9.9	1.0	1.2	-34.6	19.9	29.7	⋆⋆⋆
$\underline{\psi}_3$	-16.7	-12.9	-1.1	0.7	-1.7	-0.2	-10.2	⋆⋆⋆
$\underline{\psi}_4$	-1.7	-8.4	-178.6	2.3	-440.3	-102.6	-40.9	-20.5
$\underline{\psi}_5$	-119.3	-12.4	-1937.3	39.7	-38.1	2.7	-8.6	⋆⋆⋆

Total: 14424.88

Model II

	+	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_1$	18769.6	164.6	4.4	3.4	-20.4	12.7	-273.7	-11.2
$\underline{\psi}_2$	-1330.8	-9.6	1.0	1.2	-34.1	19.3	31.4	⋆⋆⋆
$\underline{\psi}_3$	-24.0	-12.0	-1.6	⋆⋆⋆	-0.6	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
$\underline{\psi}_4$	-1.7	-8.4	-179.0	2.3	-440.2	-102.6	-41.0	-20.4
$\underline{\psi}_5$	-119.2	-12.4	-1936.7	39.7	-38.5	2.5	-8.7	⋆⋆⋆

Total:14425.37

Model III

	+	L	O	R	L ⋆ O	L ⋆ R	O ⋆ R	L ⋆ O ⋆ R
$\underline{\psi}_1$	18712.3	160.6	4.3	2.7	-30.3	34.8	-233.2	-10.0
$\underline{\psi}_2$	-1255.8	-11.2	1.0	⋆⋆⋆	-22.2	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
$\underline{\psi}_3$	-17.6	-12.7	-1.3	0.7	-1.5	-0.1	-10.2	⋆⋆⋆
$\underline{\psi}_4$	-2.1	-8.5	-179.9	2.4	-445.3	-104.3	-36.4	-20.3
$\underline{\psi}_5$	-115.4	-11.1	-1961.2	46.0	3.1	-37.0	-9.4	⋆⋆⋆

Total: 14430.32

Of these models, Model I has the lowest discrepancy and is the model that we would select. Note that in this model the L ⋆ O ⋆ R ⋆ $\underline{\psi}_r$ contrast is not included for $r = 2,3$ and 5 but it is included for $r = 4$. Recall that $\underline{\psi}_4$ contrasts those who wish to stay with those who have a definite preference to move (to a camp either in the North or South).

The fitted probabilities under Model I (as percentages) are:

race		Black				White			
origin		North		South		North		South	
location		North	South	North	South	North	South	North	So
preference									
prefer to stay		38.0	6.0	30.0	34.0	25.0	19.0	15.0	42
prefer	North	35.9	63.5	14.7	13.8	40.4	47.6	12.2	8
to	South	8.1	11.6	30.3	29.2	10.6	9.3	50.6	33
move	Undecided	8.0	11.0	13.0	13.0	13.0	15.0	11.1	8
undecided		10.0	8.0	12.1	10.0	11.0	9.0	11.0	8

The sample proportions (as percentages)

race		Black				White			
origin		North		South		North		South	
location		North	South	North	South	North	South	North	So
preference									
prefer to stay		38.0	6.0	30.0	34.0	25.0	19.0	15.0	4
prefer to move	North	37.0	63.0	14.0	14.0	40.0	48.0	13.9	
	South	7.0	12.0	31.0	29.0	11.0	9.0	48.9	3
	Undecided	7.9	11.0	13.0	13.0	13.0	15.0	11.1	
undecided		10.1	8.0	12.1	10.0	11.0	9.0	11.1	

THE BEETLE DATA SET. The counts in this two-way table give, for each of six different dosage levels of an insecticide, the number of beetles which died and the number which survived. Here *survival* forms the row variable, with two categories: "died within 9 days" and "survived longer than 9 days", for which the model basis used $[\omega_{ir}]_{i=1,2;\ r=1,2,}$ is the Hadamard basis H_2 . *Dosage* is the column variable with six categories for which the standard orthonormal polynomial basis of order six is used. This is denoted by $[\omega_{jc}]_{j=1,\ldots,6;\ c=1,\ldots,6}$. The number of beetles subjected to each of the dosage levels is regarded as fixed. In this application we wish to model the conditional probabilities of survival for each of the given dosage levels. The models considered have the form

$$M_{ij}(\theta) = \exp\left(\frac{1}{\sqrt{2}}\sum_{c=1}^{6}\omega_{jc}\theta_{1c} + \sum_{c\in C^*}\psi_{i2}\omega_{jc}\theta_{2c}\right) \quad \text{for } i = 1,2;\ j = 1,\ldots,6$$

where $M_{+(j)}(\theta) = 1$ for $j = 1,\ldots,6$ and $C^* \subseteq \{1,\ldots,6\}$. The column effect parameters θ_{1c} for $c = 1,\ldots,6$ must be included in all models because the column totals are fixed.

There are twelve possible parameters:

		dosage (D)					
		$\underline{\omega}_1$	$\underline{\omega}_2$	$\underline{\omega}_3$	$\underline{\omega}_4$	$\underline{\omega}_5$	$\underline{\omega}_6$
	$\underline{\psi}_1$	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}
		$(+)$	(x)	(x^2)	(x^3)	(x^4)	(x^5)
survival							
(S)							
	$\underline{\psi}_2$	θ_{21}	θ_{22}	θ_{23}	θ_{24}	θ_{25}	θ_{26}
		(S)	$(x \star S)$	$(x^2 \star S)$	$(x^3 \star S)$	$(x^4 \star S)$	$(x^5 \star S)$

Since the parameters in the first row must be included in the model, attention is focussed on the parameters in the second row. The contribution tables of all the models fitted as part of the model selection are given below.

Contribution tables

Model I

213.28	0.00	0.01	0.01	0.01	0.02	
-9.69	-6.11	-0.95	-0.06	0.93	0.50	Total: 197.96

Model II

212.97	0.00	0.01	0.00	0.01	0.00	
-9.68	-6.12	-0.93	-0.07	$\star\star\star$	$\star\star\star$	Total: 196.18

Model III

212.42	0.01	0.00	0.02	0.00	0.00	
-9.68	-6.11	-0.99	$\star\star\star$	$\star\star\star$	$\star\star\star$	Total: 195.66

Model IV

211.61	-0.00	0.02	0.00	0.00	0.00	
-9.73	-6.26	$\star\star\star$	$\star\star\star$	$\star\star\star$	$\star\star\star$	Total: 195.60

There is a steady decrease in the cross-validated discrepancy as the number of parameters are excluded. Of the optional parameters the selected model (Model IV) contains only a constant term and the linear contrast parameter. We note also that the contribution of each parameter remains reasonably constant across the different models.

Once again the selected loglinear model contains the same parameters as the selected linear model. The two sets of fitted probabilities are also very similar.

The survival rates (as percentages)

	dose					
	12.08	14.49	16.31	18.31	20.44	22.36
sample	40.00	57.14	66.00	60.00	66.00	67.4
optimal linear model	46.1	53.2	57.9	62.1	66.8	70.4
optimal loglinear model	47.9	52.6	57.3	61.9	66.3	70.4

THE ESKIMO DATA. This data set involves the presence or absence of *torus mandibularis* by age (six categories) for three Eskimo populations. Let $M_{ij(k)}(\theta)$ denote the modelled probability that a randomly selected member of the kth population $(k = 1,2,3)$ falls into the ith incidence category $(i = 1,2)$ and the jth age category $(j = 1,\ldots,6)$. The models considered will be of the form

$$M_{ij(k)}(\theta) = \exp\left(\frac{1}{\sqrt{2(6)}}\sum_{\ell=1}^{3}\theta_{11\ell} + \sum_{(r,c,\ell)\in Q^*}\psi_{ir}\omega_{jc}\phi_{k\ell}\theta_{rc\ell}\right)$$

where

$$Q^* \subseteq \{(r,c,\ell) \neq (1,1,\ell) \ : \ r = 1,2;\ c = 1,\ldots 6;\ \ell = 1,2,3\}$$

and

$$M_{++(k)}(\theta) = 1 \quad \text{for} \quad k = 1,2,3.$$

The basis $[\psi_{ir}]_{i=1,2;\ r=1,2}$ used for the presence/absence variable is H_2 . For the age variable the basis, $[\omega_{jc}]_{j=1,\ldots,6;\ c=1,\ldots,6}$, used is the standard orthonormal polynomial basis of order 6, while for the three populations the basis

$[\phi_{k\ell}]_{k=1,2,3;\ \ell=1,2,3}$ used is

$$\begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} & 0 \end{pmatrix}$$

a structure which accommodates the expected similarities between the first two populations.

The contribution table (saturated model)

			populations	
incidence	age	+	$\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$
	+	1547.6	22.4	-1.7
	x	-100.5	-16.3	5.7
	x^2	5.3	3.2	1.3
+	x^3	-1.7	-0.1	0.1
	x^4	-0.4	1.5	-2.0
	x^5	0.7	1.1	0.8
	+	-8.3	5.0	1.7
	x	-189.9	-29.5	-0.9
	x^2	-9.5	-11.2	0.2
incidence	x^3	1.2	-1.5	0.7
	x^4	-0.4	0.2	1.1
	x^5	-0.6	1.9	0.7

Total: 1227.1

This table suggests a number of quasi-hierarchical models which might be worth further investigation. Since the inclusion of any of the parameters in the lower half of the table (involving the incidence contrast) necessitates the inclusion of the corresponding parameter in the upper half (where *incidence* is averaged over),

attention is focussed on the parameters in the lower half. Looking at the three columns in the lower half of the contribution table one can see that the highest order polynomial terms which make negative contributions are, respectively, x^2, x^3 (only just) and x. On the basis of this preliminary selection a number of (quasi-heirarchical) models were fitted and an optimal found. Its contribution table is given below:

The contribution table (Model I)

incidence	age	+	populations $\begin{pmatrix}1\\1\\-2\end{pmatrix}$	$\begin{pmatrix}1\\-1\\0\end{pmatrix}$
	+	1553.7	24.7	-10.5
	x	-117.3	-20.3	4.2
	x^2	6.8	3.8	⋆⋆⋆
+	x^3	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	x^4	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	x^5	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	+	-5.7	5.0	1.8
	x	-206.3	-34.5	1.8
	x^2	-3.2	-8.3	⋆⋆⋆
incidence	x^3	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	x^4	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	x^5	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆

Total: 1195.77

Remarks. 1. Only the lower order polynomials are included – quadratics being the highest order.

2. Many of the contributions in the upper half of the table are positive, but these parameters have to be included in the model since the corresponding parameters in the lower half are included and the model is to be quasi-hierarchical. The fact that these contributions are positive would indicate that there is very little difference

between the three population's age structure, and that differences only appear when the incidence variable is brought into play.

3. More parameters involving the contrast of the first two populations with the second are included than those contrasting the first two populations and most of the latter make positive contributions to the discrepancy. This indicates that the incidence by age proportions in the first two populations are similar, but that there are differences in these proportions between the first two populations and the last population. The same conclusion was reached when linear models were fitted. In fact discarding all the parameters in the third column yields a model (Model II) whose discrepancy is only slightly higher than that of Model I.

The contribution table (Model II)

			populations	
incidence	age	+	$\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$	$\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$
	+	1542.4	23.1	-1.9
	x	-112.7	-18.4	***
	x^2	6.6	3.7	***
+	x^3	***	***	***
	x^4	***	***	***
	x^5	***	***	***
	+	-4.0	52.	***
	x	-200.0	-32.7	***
	x^2	-2.5	-7.9	***
incidence	x^3	***	***	***
	x^4	***	***	***
	x^5	***	***	***

Total: 1200.89 •

THE VISION DATA. This example is concerned with the grade of vision individuals have in each of their eyes. As before the 4×4 table of counts is represented as a 16×1 vector with three components

$$\begin{pmatrix} U \\ D \\ L \end{pmatrix}$$

where D contains the four main diagonal cells, and U and L contain the upper and lower off-diagonal cells respectively. (See Section 3.2). The model basis used is

$$\begin{pmatrix} \underline{1}_6 & 0 & \underline{1}_6 & \Phi & \underline{1}_6 & \Phi \\ \underline{1}_4 & \Omega & -3\underline{1}_4 & 0 & 0 & 0 \\ \underline{1}_6 & 0 & \underline{1}_6 & \Phi & -\underline{1}_6 & -\Phi \end{pmatrix}$$

where

(1) Ω contains three contrast vectors for the four diagonal cells,

(2) Φ contains five contrast vectors for the six cells in either the upper or lower halves. The first two of these vectors contrasts the three "row averages" while the remaining three form contrasts within individual rows. (See Sections 3.2 and 4.2.)

The contribution table (saturated model)

constant	10696.6
	-140.3
main diagonal	-3.4
	-82.0
diagonal versus off-diagonal	-2811.1

	upper and lower off-diagonals	
	averaged over	contrasted
average	–	-0.4
row or	-142.0	-1.2
column totals	0.9	0.8
within individual	-51.1	1.0
rows or columns	-33.8	1.1
	-43.8	0.0

Total: 7391.85

The next model fitted (Model II) contains only those parameters whose contributions in the saturated model are negative.

The contribution table (Model II)

constant	10698.2
	-140.3
main diagonal	-3.4
	-82.0
diagonal versus off-diagonal	-2812.5

	upper and lower off-diagonals	
	averaged over	contrasted
average	–	-0.5
row or	-142.5	-1.0
column totals	(A) $\star\star\star$	(B) $\star\star\star$
within individual	-51.1	$\star\star\star$
rows or columns	-33.7	$\star\star\star$
	-42.3	$\star\star\star$

Total: 7388.83

Two other similar models were fitted. Both contained all of the parameters of Model II together with

(a) the parameter marked (A) above

(b) the parameters (A) and (B).

The cross-validated discrepancies of these two models were found to be 7390.0 and 7390.4 respectively, i.e. higher than the discrepancy for Model II. Noting that Model II contains parameters which model differences between the upper and lower halves and hence will not give a model of symmetry, two symmetry models were also fitted. Their cross-validated discrepancies were once again higher than that of Model II, (in fact they are even higher than that of the saturated model). The fitted probabilities under Model II are:

	grade of left eye				
grade of right eye	highest (1)	second (2)	third (3)	lowest (4)	Totals
highest (1)	25.3	3.4	2.5	1.1	32.3
second (2)	3.4	15.2	4.6	0.8	24.0
third (3)	2.5	4.6	18.0	2.7	27.8
lowest (4)	1.3	1.0	3.3	10.2	15.8
Totals	32.5	24.2	28.4	14.8	100

The selected linear model for this data set is symmetric whereas the selected loglinear model is not. This is, of course, a consequence of the fact that we are using different discrepancies. In the linear case the emphasis is on the absolute difference between probabilities, whereas in the loglinear case it is on the ratio of probabilities. It can be seen that the asymmetries in the above table occur in cells with relatively low probabilities. Thus, whereas the probabilities in symmetric pairs of cells might differ little in absolute terms, their ratio's can differ substantially.

CHAPTER 6
ROTATION INVARIANCE

This chapter is concerned with so-called cyclical categorical variables whose categories, like the days of the week, have a cyclical or circular ordering. For a cyclical variable with L categories there are, depending on which of the categories is placed first, L distinct ways in which the categories can be listed, each of which is called a *rotation.* For such a variable the modelling procedure used should, for any two rotations, fit essentially the same model – that is the two fitted models should differ only in the ordering of the categories while the fitted probabilities for each specific cell should be identical. Such modelling procedures are said to be *rotation invariant.* In this chapter we investigate the conditions under which the modelling procedure for linear basis models is rotation invariant.

The first section deals with the invariance of the procedure for a single random variable. It is found that the rotation invariance of the procedure is a function of the number of categories (L) and that for some L rotation invariance cannot be guaranteed. In Section 3 a modification of the linear modelling procedure is introduced which ensures that rotation invariance can be achieved for variables with any number of categories. Sections 2 and 4 give the corresponding extensions to multiway cross-classifications in which some or all of the variables are cyclical, together with some examples of applications. The proofs of the results in this chapter are somewhat technical and are given in Appendices B and C.

6.1 LINEAR MODELS

Consider a cyclical variable with L categories. Choose any particular listing or rotation of the categories and label them as

$$\begin{pmatrix} 1 \\ \vdots \\ L \end{pmatrix}.$$

This vector can be transformed into any of the other rotations by "rotating" its

elements the required number of times. We use the convention of bottom-to-top (rather than top-to-bottom) rotations.

Formally, an operator which will rotate the elements of a $L \times 1$ vector n times is defined by the partitioned matrix

$$R^n = \begin{pmatrix} 0_{n\times(L-n)} & I_n \\ I_{L-n} & 0_{(L-n)\times n} \end{pmatrix} \quad \text{for} \quad n = 0, \ldots, L.$$

Note that the R^n are orthogonal with

$$(R^n)^{-1} = (R^n)' = R^{L-n}.$$

Suppose that we have L categories and a $L \times L$ model basis. Let $\underline{M}_{R^n}$ $(n = 0, 1, \ldots, L-1)$ denote the model obtained when the categories and the associated cell counts are rotated n times and presented to the modelling procedure, which is understood to use the same model basis for all rotations of the categories. For the non-cyclical variables considered in Chapter 4 the vectors were chosen to contrast specific cells and one would expect the $\underline{M}_{R^n}$ to be quite different from one another. For cyclical variables on the other hand, the models obtained for two different rotations should differ only by the corresponding rotation, i.e.

$$\underline{M}_{R^n} = R^n(\underline{M}_{R^0}) \quad \text{for} \quad n = 0, \ldots, L-1.$$

Let Φ be the model basis that is used. Using the notation introduced in Section 3.3

$$R^n(\underline{M}_{R^0}) = \sum_{q \in Q_{R^0}} (\underline{\phi}_q \cdot \underline{P}) R^n \underline{\phi}_q$$

with

$$q \in Q_{R^0} \quad \text{iff} \quad (\underline{\phi}_q \cdot \underline{P})^2 / \hat{\text{var}}(\underline{\phi}_q \cdot \underline{P}) > 2,$$

while

$$\underline{M}_{R^n} = \sum_{q \in Q_{R^n}} (\underline{\phi}_q \cdot R^n \underline{P}) \underline{\phi}_q$$

with

$$q \in Q_{R^n} \quad \text{iff} \quad (\underline{\phi}_q \cdot R^n \underline{P})^2 / \hat{\text{var}}(\underline{\phi}_q \cdot R^n \underline{P}) > 2\,.$$

Theorem 1. The modelling procedure for linear basis models is rotation invariant iff the basis Φ, is such that,

$$\left.\begin{array}{l}\text{for all } n\ (n=1,\ldots,L-1) \text{ and for all } \ \underline{\phi}_q \in \Phi \\ \text{there exists a } \ \underline{\phi}_r \in \Phi \ \text{ such that } \ R^n\underline{\phi}_q = \underline{\phi}_r \ \text{ or } \ R^n\underline{\phi}_q = -\underline{\phi}_r .\end{array}\right\} \quad (1)$$

Theorem 1 allows one to concentrate on the basis used, rather than the entire modelling procedure. In view of condition (1) two definitions are made.

Definition. Two vectors $\underline{\phi}$ and $\underline{\psi}$ are defined to be $\sim$-equivalent, written $\underline{\phi} \sim \underline{\psi}$, if $\underline{\phi} = \underline{\psi}$ or $\underline{\phi} = -\underline{\psi}$.

Note that changing the sign of a vector in a model basis has no effect on the modelling procedure. The only effect that it does have is to change the sign of the associated parameter.

Definition. A model basis, Φ, (as opposed to a modelling procedure) is defined to be rotation invariant if, for all n and for all $\underline{\phi}_q \in \Phi$, there exists a $\underline{\phi}_r \in \Phi$ such that $R^n\underline{\phi}_q \sim \underline{\phi}_r$.

The rest of this section is concerned with (i) determining conditions under which rotation invariant model bases exist, and (ii) characterising the form of the bases when they exist.

As a beginning note that it follows from Theorem 1 that if $\underline{\phi}$ is in a rotation invariant model basis, then all of the $\sim$-distinct rotations of $\underline{\phi}$ must also be in the basis.

Definition. The set of all $\sim$-distinct rotations of a vector $\underline{\phi}$, is defined by $R(\underline{\phi}) = \{R^n\underline{\phi} : \ n = 0,\ldots,m-1$ where m is the least positive integer such that $R^m\underline{\phi} \sim \underline{\phi}\}$.

$R(\underline{\phi})$ is called the *rotation group generated by* $\underline{\phi}$ and is said to have *cardinality* m.

It follows that rotation invariant model bases will consist of rotation groups. Now the vector $\frac{1}{\sqrt{L}}\underline{1}_L$ generates the rotation group $\{\frac{1}{\sqrt{L}}\underline{1}_L\}$ of cardinality one. Since this vector must be in all model bases it follows that any other rotation group apearing in a rotation invariant model basis can have cardinality at most $L-1$. The following corollary follows immediately from Theorem 1.

Corollary 2. All rotation invariant model bases consist of two or more rotation groups, whose cardinalities will sum to L, and each of whose cardinalities will be at most $L-1$. •

Example. Let $L=4$ and put

$$\underline{\phi}=\frac{1}{2}\begin{pmatrix}1\\-1\\1\\-1\end{pmatrix},\underline{\psi}=\frac{1}{2}\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix}.$$

The cardinalities of $\underline{\phi}$ and $\underline{\psi}$ are one and two respectively, and

$$\left(R(\tfrac{1}{2}\underline{1}_4)\,,\,R(\underline{\phi})\,,\,R(\underline{\psi})\right)$$

$$=\frac{1}{2}\begin{pmatrix}1&1&1&-1\\1&-1&1&1\\1&1&-1&1\\1&-1&-1&-1\end{pmatrix}$$

is a rotation invariant model basis which consists of three rotation groups. •

Having established that rotation invariant model bases consist of rotation groups we now investigate the form of these rotation groups, or equivalently the form of their generators. For a vector, $\underline{\phi}$, to generate a rotation group of cardinality n, $\underline{\phi}$ must satisfy at least $R^n\underline{\phi}\sim\underline{\phi}$.

Theorem 3. All vectors satisfying $R^n\underline{\phi} \sim \underline{\phi}$ for some n $(1 \leq n \leq L-1)$ are of one of the two forms

$$\underline{1}_d \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix} \tag{2}$$

or

$$\underline{\Lambda}_d \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix} \tag{3}$$

where

(i) in (2) d is any divisor of L, while in (3) d must be an *even* divisor of L,

(ii)

$$\underline{\Lambda}_d = \underline{1}_{d/2} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

(iii) $a_i \in R$ for $i = 1, \ldots, L/d$. •

This theorem provides a partial characterisation of rotation group generators. However the orthogonality requirements of a model basis are yet to be considered. In terms of the constituent rotation groups the orthogonality requirements are:

(i) each rotation group must be orthogonal, (i.e. all the vectors within a rotation group must be orthogonal to one another), and

(ii) all rotation groups must be orthogonal to each other, (i.e. each vector in each rotation group must be orthogonal to all of the vectors in each other rotation group).

Consider the orthogonality of individual rotation groups first.

Lemma 4. (a) Let $\underline{\phi}$ have the form (2) with $L/d \geq 2$. Then $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} a_{L/d} & a_1 & a_2 & \cdots & a_{(L/d)-1} \\ a_{(L/d)-1} & a_{L/d} & a_1 & \cdots & a_{(L/d)-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_2 & a_3 & a_4 & \cdots & a_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L/d} \end{pmatrix} = \underline{0} \tag{4}$$

(b) Let $\underline{\phi}$ have the form (3) with $L/d \geq 2$. Then $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} -a_{L/d} & a_1 & a_2 & \cdots & a_{(L/d)-1} \\ -a_{(L/d)-1} & -a_{L/d} & a_1 & \cdots & a_{(L/d)-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_2 & -a_3 & -a_4 & \cdots & a_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L/d} \end{pmatrix} = \underline{0} \qquad (5)$$

(c) Both (4) and (5) are consistent systems. •

Before considering orthogonality between pairs of rotation groups in general, we look at the orthogonality of the rotation groups generated by vectors of the form (2) and (3), with the rotation group $\{\frac{1}{\sqrt{L}}\underline{1}\}$. Clearly the elements of vectors of the form (3) will sum to zero; so that these rotation groups will be orthogonal to $\{\frac{1}{\sqrt{L}}\underline{1}\}$. The same does not hold for generators of the form (2), and one may ask whether the group may be made orthogonal to $\frac{1}{\sqrt{L}}\underline{1}$ by a suitable choice of the elements $a_1, \ldots, a_{L/d}$. The next proposition answers this question.

Proposition 5. Let $\underline{\phi}$ be a non-zero vector with the form (2). $R(\underline{\phi})$ cannot both be orthogonal to $\{\frac{1}{\sqrt{L}}\underline{1}\}$ and have its vectors orthogonal to one another. •

This proposition implies that in a rotation invariant model basis all the rotation group generators besides $\frac{1}{\sqrt{L}}\underline{1}_L$ must be of the form (3) where the divisor involved must be even. This has an interesting consequence. Suppose that L is odd, then L has no even divisors and hence no rotation groups of the form (3) exist. Thus there are no rotation groups, besides $\frac{1}{\sqrt{L}}\underline{1}_L$, which can be used to form a model basis. Thus no rotation invariant model basis can exist for odd L. The problem of the non-existence of rotation invariant model bases is considered in section 3. For the time being we concentrate on the case where L is even.

We have established that besides $\frac{1}{\sqrt{L}}\underline{1}_L$ all rotation group generators appearing in a model basis are of the form (3), and that in order that the vectors within each rotation group be orthogonal to one another the condition (5) must be satisfied.

The next proposition is concerned with the orthogonality of one rotation group to another. Before giving the proposition we define the *greatest common divisor* (highest common factor) of two natural numbers d_1 and d_2 by

$$gcd(d_1, d_2) = \max\{m : \tfrac{d_1}{m} = a, \tfrac{d_2}{m} = b \text{ for some natural numbers } a \text{ and } b\} .$$

Proposition 6. Two rotation groups of cardinalities $\frac{L}{d_1}$ and $\frac{L}{d_2}$ generated by vectors of the form (3) and satisfying the orthogonality restrictions given by (5), will be orthogonal to one another iff

$$\frac{d_1 + d_2}{gcd(d_1, d_2)} \quad \text{is odd} \tag{6}$$

with no further restrictions on the elements of the generators. •

This result is rather surprising in that the condition for orthogonality between two rotation groups does not involve restrictions on the elements of the two generators but only a restriction on their cardinalities. However, that (6) must hold for all pairs of rotation groups which are to appear in the same basis, turns out to be an extremely restrictive condition. Note that in particular it rules out the possibility of having two (or more) rotation groups of the same cardinality (≥ 2) in a model basis. In fact, as will be shown, this condition means that for many (even) L there are less than L vectors which may be simultaneously included in a rotation invariant model basis – which means that no such basis can exist for that L.

Proposition 7. The maximum number of vectors available for simultaneous inclusion in a rotation invariant model basis is less than or equal to L with equality holding only if $L = 2^m$ for some natural number m. •

This proposition implies that for a rotation invariant model basis to exist it is necessary that L be a power of 2. Sufficiency is easy to establish. If $L = 2^m$ then the permissible cardinalities for generators of the form (3) are

$$2^s \quad \text{for} \quad s = 0, 1, \ldots, m-1.$$

Thus, including $\frac{1}{\sqrt{L}}\underline{1}_L$, the total number of vectors available is

$$1 + \sum_{s=0}^{m-1} 2^s = 2^m$$

and a basis can be formed by using generators of each of the above cardinalities.

In order to write down the general form of the associated rotation invariant model basis, we will use $\underline{r}_{2^s}$ to denote a vector which has the form

$$\underline{\Lambda}_{2^{m-s}} \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{2^s} \end{pmatrix}.$$

All rotation invariant model bases must then (barring normalising factors) be of the form

$$\left(\frac{1}{\sqrt{L}}\underline{1}_L; \quad R(\underline{r}_{2^{m-1}}); \quad R(\underline{r}_{2^{m-2}}); \ldots; R(\underline{r}_{2^0})\right) \tag{7}$$

where each generator of cardinality 2 or more satisfies (5) of Lemma 4(b). The next proposition is a refinement of Lemma 4(b).

Proposition 8. Let $L = 2^m$ and let

$$\underline{\phi} = \underline{\Lambda}_{2^{m-s}} \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{2^s} \end{pmatrix} \quad \text{for some } s, \quad 0 \leq s \leq m-1.$$

(a) For $s = 0$ and 1; $R(\underline{\phi})$ is orthogonal.

(b) For $s = 2, \ldots, m-1$; $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} a_{2^s} & a_1 & a_2 & \cdots & a_{2^s-1} \\ -a_{2^s-1} & -a_{2^s} & a_1 & \cdots & a_{2^s-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{\frac{1}{2}2^s+2} & -a_{\frac{1}{2}2^s+3} & -a_{\frac{1}{2}2^s+4} & \cdots & -a_{\frac{1}{2}2^s+1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_2^s \end{pmatrix} = \underline{0} \tag{8}$$

which is a consistent system. •

The above results can be summarised as follows.

Theorem 9. Rotation invariant model bases can exist only when $L = 2^m$ for some natural number m. If $L = 2^m$, then a rotation invariant model basis must be of the form (7) where each of the rotation group generators must be normalised, and those of cardinality four or more must satisfy the restrictions as given by (8).

•

This theorem entirely characterises the form of rotation invariant model bases. The next section investigates applications of and extensions to this result; while in the following section a modification to the modelling procedure is proposed which allows rotation invariance to be guaranteed for all cyclical variables no matter how many categories they have.

•

6.2 APPLICATIONS AND EXTENSIONS

Having derived the theory for modelling single cyclical variables (with 2^m categories) we will in this section

(1) consider in more detail the construction of rotation invariant model bases,

(2) fit a model, using real data, to the classification of a cyclical variable,

(3) extend the theory to multiway cross-classifications involving cyclical variables, and

(4) fit a model to a cross-classification which has one cyclical and one non-cyclical variable.

Choosing bases. Consider fitting a linear basis model to the classification of a cyclical variable which has 2^m categories. The form of the basis that must be used is determined by Theorem 1.9. For example the form (barring normalising factors) of the smallest three rotation invariant model bases are given by Theorem 1.9 as:
$L = 2$

$$(\underline{1}_2;\ \underline{\Lambda}_2)$$

$L = 4$

$$\left(\underline{1}_4;\ R\left(\underline{\Lambda}_2 \otimes \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right);\ \underline{\Lambda}_4\right)$$

$L = 8$

$$\left(\underline{1}_8;\ R\left(\underline{\Lambda}_2 \otimes \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}\right);\ R\left(\underline{\Lambda}_4 \otimes \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right);\ \underline{\Lambda}_8\right)$$

where the b_i have to satisfy

$$(-b_4, b_1, b_2, b_3) \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = 0.$$

We are only free to choose the elements of the generators. Since the elements in the generators have to be normalised we can, without loss of generality, take any

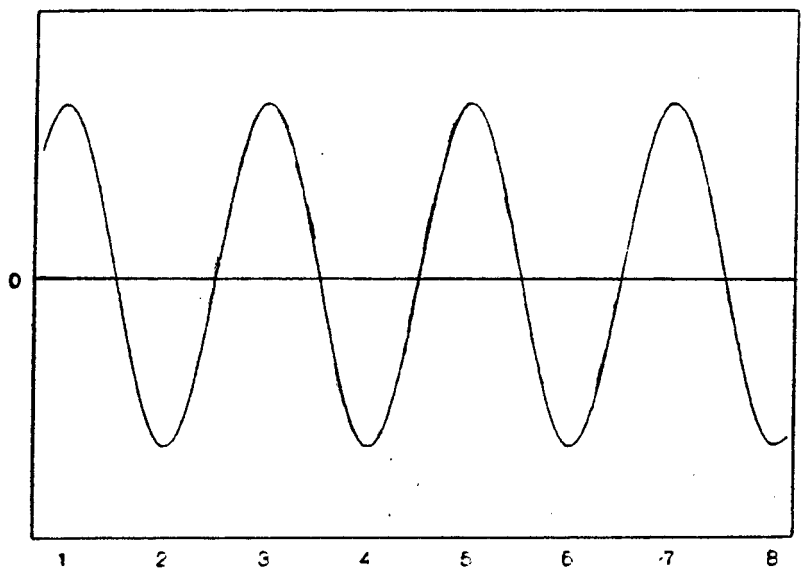
one of the elements in a generator to be 1. We will adopt the convention that this element will be the topmost in the generator. Some examples of generators are then

$$\underline{\Lambda}_L;\ \underline{\Lambda}_{L/2} \otimes \begin{pmatrix} 1 \\ a \end{pmatrix};\ \underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix};\cdots$$

We consider now the choice of the elements of the generators. As a beginning note that plotting and joining the points

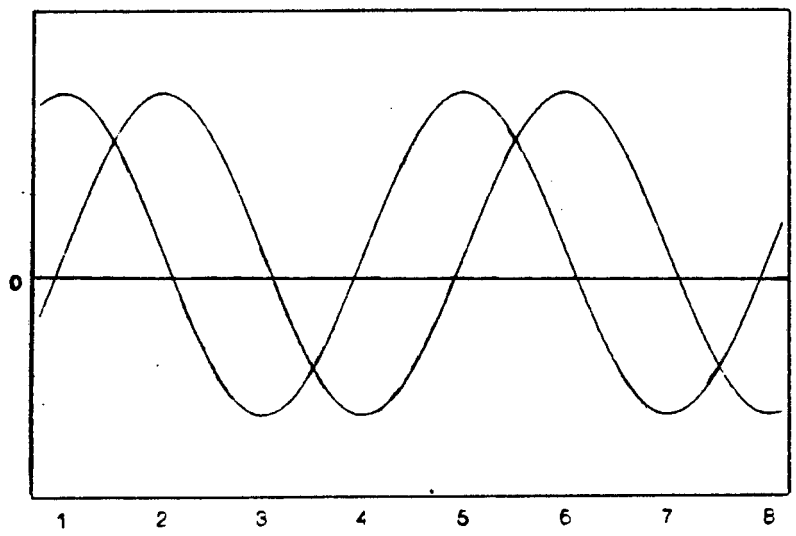
$$(i, \Lambda_{iL})_{i=1,\ldots,L} = (i, (-1)^{i+1})_{i=1,\ldots,L}$$

gives a sinusiodal wave.



(1)

We will endeavour to construct all generators, $\underline{\phi}$, so that a plot of $(i, \phi_i)_{i=1,\ldots,L}$ produces such a wave. Generally the frequency of the wave will decrease (the wave becomes smoother) as the cardinality of the generator increases. For example, by choosing $a = 0$ in $(\underline{\Lambda}_{L/2} \otimes \binom{1}{a})$ the plots for the two vectors in the rotation group are again sinusoidal waves; whose frequency is half that of the wave in (1).



(2)

We consider next the choice of the b_i in $\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}$, which have to satisfy the orthogonality constraint:

$$-b_3 + b_1 + b_1 b_2 + b_2 b_3 = 0\,. \tag{3}$$

In order to continue the pattern we would like to choose the b_i so as to obtain a wave with half the frequency of those in (2). The natural choice of values are those from the standard sine-wave, namely

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix}. \tag{4}$$

However this choice does not satisfy the orthogonality restriction (3). In fact it has not been found possible to produce b_i which simultaneously satisfy (3) and produce a wave of the required form. The generators given below satisfy (3) but do not produce sinusoidal waves. They have been used to fit a number of data sets

and in all cases produced models with similar estimated expected discrepancies and fitted probabilities.

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \tag{5}$$

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{6}$$

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{7}$$

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ -1+\sqrt{2} \\ -1+\sqrt{2} \\ 1 \end{pmatrix} \tag{8}$$

$$\underline{\Lambda}_{L/4} \otimes \begin{pmatrix} 1 \\ -1-\sqrt{2} \\ -1-\sqrt{2} \\ 1 \end{pmatrix} \tag{9}$$

Of these generators (5) represents what may be considered the smoothest wave and is the generator that is used in the applications discussed below.

Parameter and criterion tables. In the illustrative examples given we will use parameter and criterion tables as introduced previously, but with one difference. When a cyclical variable is involved each vector in the basis no longer has associated with it a unique (estimated) parameter value since the parameter values depend on the particular rotation of the categories that is used. However, for each rotation group the set containing the absolute values of the associated parameters, remains the same for all rotations, i.e. all that happens is that with each different rotation of the cells the order (and possibly the signs) of the parameters are altered, but the actual absolute *values* are unaffected. Consequently in giving the parameter table we simply give, for each generator, the set of absolute values of the associated parameters.

Each parameter estimate still has, of course, a unique (estimated) standard deviation and contribution to the expected discrepancy. These are included in the relevant tables.

Example. The data set introduced here will be used repeatedly in this chapter. (In fact it provided the motivation for developing a rotation invariant modelling procedure.) The data consists of hourly measurements of wind direction and speed made 10 metres above the ground at the meterological station attached to the Koeberg Nuclear Power Station. Measurements are made every 5 minutes and the (direction, speed) value given for a particular hour is the average of the measurements made during the preceeding hour. The data are being collected primarily for constructing models for the short-term prediction of the direction and speed of the wind in the event of a radioactive leak in the plant. We will only use subsets of the data and will only construct models that will assist in illustrating the theory that has been developed. Consider first the eight direction segments and the data for a single month. The sample counts for December 1979 are shown:

N	NW	W	SW	S	SE	E	NE
171	118	66	85	84	62	50	108

Using the basis (which in non-normalised form is)

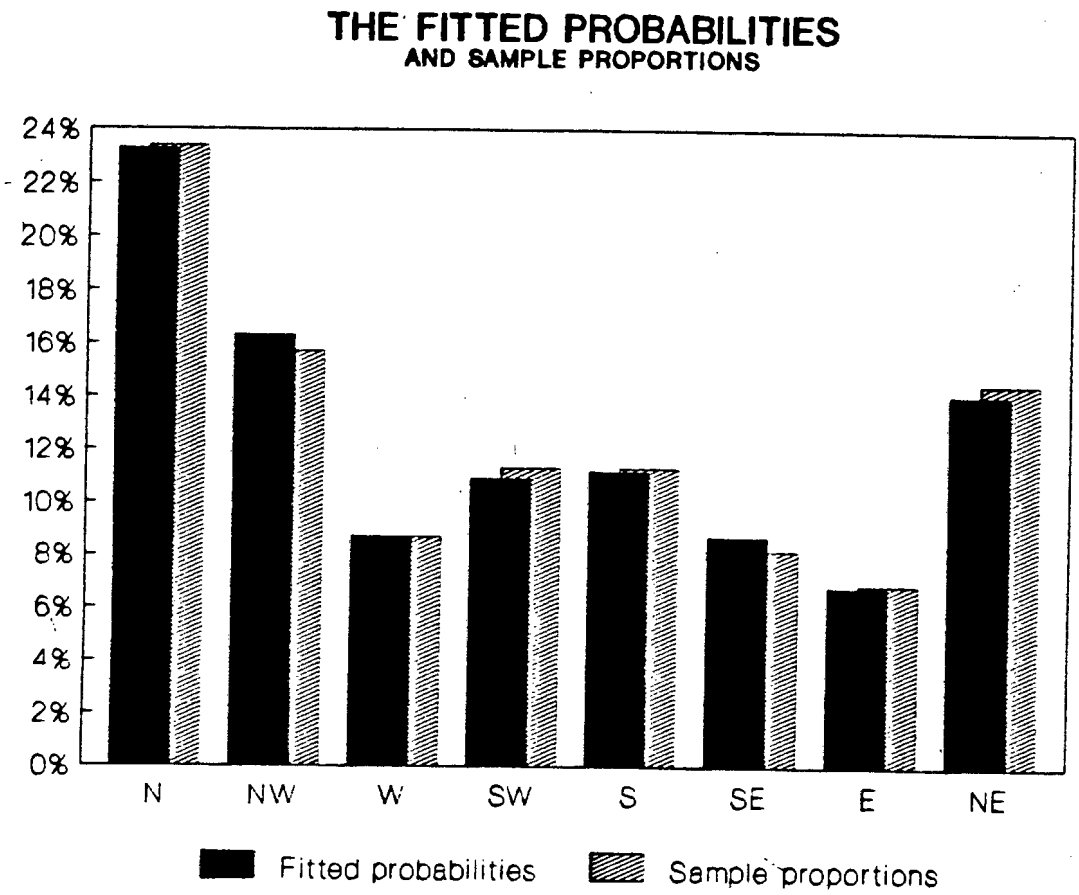
$$\left(\underline{1}_8;\ R\left(\underline{\Lambda}_2 \otimes \begin{pmatrix}1\\0\\1\\0\end{pmatrix}\right);\ R\left(\underline{\Lambda}_4 \otimes \begin{pmatrix}1\\0\end{pmatrix}\right);\ \underline{\Lambda}_8\right)$$

yields the following parameter and criterion tables:

The parameter and criterion tables (all entries $\times 10^3$)

	$\underline{1}_8$	$R\left(\underline{\Lambda}_2 \otimes \begin{pmatrix}1\\0\\1\\0\end{pmatrix}\right)$				$R\left(\underline{\Lambda}_4 \otimes \binom{1}{0}\right)$		$\underline{\Lambda}_8$
parameter	353.6	69.9	21.5	51.1	53.8	95.4	9.4	1.9
std.deviation	0	12.8	12.9	12.9	12.8	12.5	12.9	13.0
contribution	-125.0	-4.6	-0.1	-2.3	-2.6	-8.8	0.2	0.3

The selection criterion indicates that two of the seven parameters should not be included in the final fitted model. These parameters are those associated with the higher frequency waves.



More than one variable. So far we have only considered a single cyclical variable. In practice one may have to deal with multiway cross-classifications, in which some of the variables may be cyclical and some not.

Consider initially two random variables X and Y; with R and C categories respectively and let Ψ and Ω be their respective model bases. The linear modelling procedure is then characterised by the fitted model which it produces, namely

$$M_{ij}(\hat{\theta}) = \sum_{(r,c)\in A} \psi_{ir}\omega_{jc}\hat{\theta}_{rc}$$

where

(i) $(r,c) \in A$ iff $\hat{\theta}_{rc}^2 - 2\hat{\text{var}}\,\hat{\theta}_{rc} < 0$ for $r = 1,\ldots,R;\ c = 1,\ldots,C$

(ii) $$\hat{\theta}_{rc} = \sum_i \sum_j \psi_{ir}\omega_{jc}P_{ij}$$

(iii) $$\hat{\text{var}}\,\hat{\theta}_{rc} = \sum_j \omega_{jc}^2 \frac{1}{(n_{+j}-1)}\left\{\sum_i \psi_{ir}^2 P_{ij} - (\sum_i \psi_{ir}P_{ij})^2\right\}$$

and where $M_{ij}(\hat{\theta})$ and P_{ij} may refer to multinomial or product-multinomial probabilities.

Suppose that X is a cyclical variable. It is not difficult to see that if Ψ is a rotation invariant basis that the modelling procedure is invariant with respect to rotations of the categories of X.

The same clearly holds for Y, and for X and Y simultaneously. That is, if Ψ and Ω are both rotation invariant then the modelling procedure is invariant to rotations of the categories of either or both variables.

The above considerations generalise to cross-classifications involving more than two variables. No matter how many variables appear in a cross-classification, the modelling procedure will be invariant to rotations of the categories of any variable for which a rotation invariant model basis is used.

Example. Consider the Koeberg wind data discussed above where now the (non-cyclical) variable wind speed is introduced. Wind speeds have been divided into three categories.

	speed			
direction	0-3.9	4-7.9	8+	totals
N	20	82	69	171
NW	22	77	19	118
W	12	49	5	66
SW	30	51	4	85
S	41	38	5	84
SE	36	24	2	62
E	28	20	2	50
NE	24	68	16	108
	213	409	122	744

Both variables are treated as response variables, so that the operating model is multinomial. The joint basis can now be constructed from two bases, one for each variable. For direction we can use the same basis as was used in the previous example namely,

$$\left(\underline{1}_8;\quad R\left(\underline{\Lambda}_2 \otimes \begin{pmatrix}1\\0\\1\\0\end{pmatrix}\right);\quad R\left(\underline{\Lambda}_4 \times \begin{pmatrix}1\\0\end{pmatrix}\right);\quad \underline{\Lambda}_8\right)$$

which will be written as

$$(\underline{1}_8; R(\underline{a}); R(\underline{b}); \underline{\Lambda}_8).$$

For wind speed, for example, an orthogonal polynomial basis (of order three) can be used, which in non-normalised form is

$$\begin{pmatrix}1 & -1 & 1\\ 1 & 0 & -2\\ 1 & 1 & 1\end{pmatrix}.$$

The resulting joint basis leads to the following parameter and criterion tables.

The parameter table

direction	speed $\begin{pmatrix}1\\1\\1\end{pmatrix}$	$\begin{pmatrix}-1\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\-2\\1\end{pmatrix}$
	204.1	30.6	93.7
$\underline{1}_8$	(0.0)	(6.1)	(7.9)
	40.0	49.4	31.8
$R(\underline{a})$	(7.3)	(6.2)	(8.3)
	12.8	6.2	20.6
	(7.5)	(5.9)	(8.8)
	27.6	31.4	7.1
	(7.4)	(6.3)	(8.4)
	30.6	23.3	35.9
	(7.4)	(5.8)	(8.7)
	53.9	21.9	3.8
$R(\underline{b})$	(7.2)	(6.4)	(8.4)
	5.0	1.4	11.2
	(7.5)	(5.9)	(8.8)
	0.6	17.1	17.6
$\underline{\Lambda}_8$	(7.5)	(6.1)	(8.6)

The criterion table

direction	speed $\begin{pmatrix}1\\1\\1\end{pmatrix}$	$\begin{pmatrix}-1\\0\\1\end{pmatrix}$	$\begin{pmatrix}1\\-2\\1\end{pmatrix}$
$\underline{1}_8$	-41.7	-0.9	-8.7
$R(\underline{a})$	-1.5	-2.4	-0.9
	-0.1	0.0	-0.3
	-0.6	-0.9	0.1
	-0.8	-0.5	-1.1
$R(\underline{b})$	-2.8	-0.4	0.1
	0.1	0.1	0.0
$\underline{\Lambda}_8$	0.1	-0.2	-0.2

Of the 23 possible parameters the selection criterion indicates that 7 should not be included in the fitted model. Note that four of the six parameters associated with the rotations of $\underline{\Lambda}_4 \otimes \binom{1}{0}$ are marked for exclusion.

The fitted probabilities (as percentages)

direction	speed 0-3.9	4-7.9	8+	
N	2.6	11.2	9.2	23.0
NW	3.0	10.9	2.3	16.2
W	1.7	6.5	0.8	9.0
SW	4.4	6.2	0.3	10.9
S	5.7	4.7	0.9	11.3
SE	4.5	3.8	0.4	8.7
E	3.6	3.1	0.1	6.8
NE	3.1	8.5	2.4	14.0
	28.6	54.9	16.5	100

The sample proportions (as percentages)

	speed			
direction	0-3.9	4-7.9	8+	
N	2.7	11.0	9.3	23.0
NW	3.0	10.3	2.6	15.9
W	1.6	6.6	0.7	8.9
SW	4.0	6.9	0.5	11.4
S	5.5	5.1	0.7	11.3
SE	4.8	3.2	0.3	8.3
E	3.8	2.7	0.3	6.8
NE	3.2	9.1	2.2	14.5
	28.6	54.9	16.5	100

6.3 PAIRWISE LINEAR MODELS

For many applications in which we would wish to fit rotation invariant models - the number of categories, L, is not a power of 2. For example seasonal data of this type often have $L = 12$ (monthly counts), $L = 52$ (counts in pentads) or even $L = 365$ (daily counts). In this section we consider a modification of the notion of rotation invariance which preserves most of the desirable properties and which can be applied for an arbitrary number of categories.

To introduce the modification, consider grouping the vectors in some model basis Φ into pairs, as follows:

$$\left(\frac{1}{\sqrt{L}}\underline{1}_L,\ (\underline{\phi}_2,\underline{\phi}_3),(\underline{\phi}_4,\underline{\phi}_5),\ldots,(\underline{\phi}_{L-1},\underline{\phi}_L)\right) \quad \text{for odd } L$$

and

$$\left(\frac{1}{\sqrt{L}}\underline{1}_L,\ (\underline{\phi}_2,\underline{\phi}_3),(\underline{\phi}_4,\underline{\phi}_5),\ldots,(\underline{\phi}_{L-2},\underline{\phi}_{L-1}),\underline{\phi}_L\right) \quad \text{for even } L.$$

One has to be careful about the definition of these pairs of vectors. Since changing the order of the elements within the pair, and/or changing the sign of either (or both) element(s) does not affect the modelling procedure, we will define two pairs $(\underline{\phi}_1,\underline{\phi}_2)$ and $(\underline{\psi}_1,\underline{\psi}_2)$ to be non-distinct if

$$(1) \quad \underline{\phi}_1 = \pm\underline{\psi}_1, \quad \underline{\phi}_2 = \pm\underline{\psi}_2$$

or

$$(2) \quad \underline{\phi}_1 = \pm\underline{\psi}_2, \quad \underline{\phi}_2 = \pm\underline{\psi}_1.$$

The second complicating factor is the difference between odd and even L. We will begin by restricting attention to the more simple of the two cases, namely that of odd L. Later it will be shown that the theory developed for odd L is easily extended to incorporate even L.

As before, the saturated model can be written as a linear combination of all the vectors in Φ; but we now maintain the pairing, i.e. the saturated model is

$$\frac{1}{L}\underline{1}_L + \sum_{q=1}^{\frac{1}{2}(L-1)} \left\{(\underline{\phi}_{2q}\cdot\underline{P})\underline{\phi}_{2q} + (\underline{\phi}_{2q+1}\cdot\underline{P})\underline{\phi}_{2q+1}\right\}.$$

The modification to the selection criterion is that we now include or exclude the terms in pairs; never one element of a pair without the other. The criterion used to determine whether each pair should be included is based on the contributions to the expected discrepancy of both parameters in the pair. In fact the criterion is:

include the pair only if the *sum* of the individual contributions from each of the elements in a pair is negative.

More formally, the modelling procedure under consideration is that which produces, for a given rotation of the cell probabilities the fitted model

$$\frac{1}{\sqrt{L}}\underline{1}_L + \sum_{q\in Q}\left\{(\underline{\phi}_{2q}\cdot\underline{P})\underline{\phi}_{2q} + (\underline{\phi}_{2q+1}\cdot\underline{P})\underline{\phi}_{2q+1}\right\}$$

where $Q \subseteq \{1,2,\ldots,\frac{1}{2}(L-1)\}$ and $q \in Q$ iff

$$\left[(\underline{\phi}_{2q}\cdot\underline{P})^2 - 2\,\hat{\text{var}}(\underline{\phi}_{2q}\cdot\underline{P})\right] + \left[(\underline{\phi}_{2q+1}\cdot\underline{P})^2 - 2\,\hat{\text{var}}(\underline{\phi}_{2q+1}\cdot\underline{P})\right] < 0.$$

This procedure will be called the *pairwise modelling procedure.*

Theorem 1. The pairwise modelling procedure is rotation invariant iff the basis Φ , is such that

for all n $(1 \le n \le L)$ and for all $q \in \{1,\ldots,\frac{1}{2}(L-1)\}$ there exists an $r \in \{1,\ldots,\frac{1}{2}(L-1)\}$ such that

$$\phi_{i,2r}\phi_{j,2r} + \phi_{i,2r+1}\phi_{j,2r+1}$$
$$= (R^n\underline{\phi}_{2q})_i(R^n\underline{\phi}_{2q})_j + (R^n\underline{\phi}_{2q+1})_i(R^n\underline{\phi}_{2q+1})_j \quad \text{for all } i,j$$

where $(R^n\underline{\phi})_i$ denotes the ith element of the vector $R^n\underline{\phi}$. •

As the analog to the $\sim$ −equivalence of the previous section define $\approx$ −equivalence by

$$(\underline{\phi}_1,\underline{\phi}_2) \approx (\underline{\psi}_1,\underline{\psi}_2)$$

if

$$\phi_{i1}\phi_{j1} + \phi_{i2}\phi_{j2} = \psi_{i1}\psi_{j1} + \psi_{i2}\psi_{j2} \quad \text{for all } i,j. \tag{1}$$

Note that exchanging a pair of vectors in a basis with a pair which is $\approx$ −equivalent will not affect the model which is fitted by the pairwise fitting procedure. To see this multiply (1) through by P_i and then sum over all i to get

$$(\underline{\phi}_1 \cdot \underline{P})\underline{\phi}_1 + (\underline{\phi}_2 \cdot \underline{P})\underline{\phi}_2 = (\underline{\psi}_1 \cdot \underline{P})\underline{\psi}_1 + (\underline{\psi}_2 \cdot \underline{P})\underline{\psi}_2.$$

Theorem 1 can be restated in terms of $\approx$ −equivalence among the vector pairs in the basis as: the pairwise modelling procedure is rotation invariant iff the basis used, Φ, is such that

$$\left.\begin{array}{l}\text{for all } n \text{ and for all } (\underline{\phi}_{2q}, \underline{\phi}_{2q+1}) \text{ there exists } (\underline{\phi}_{2r}, \underline{\phi}_{2r+1}) \\ \text{such that } (\underline{\phi}_{2r}, \underline{\phi}_{2r+1}) \approx (R^n\underline{\phi}_{2q}, R^n\underline{\phi}_{2q+1}).\end{array}\right\} \quad (2)$$

Condition (2) can be simplified. It is not difficult to show that two distinct vector pairs which are $\approx$ −equivalent cannot appear in the same model basis (as all four vectors cannot each be orthogonal to one another). Hence for given n and $(\underline{\phi}_{2q}, \underline{\phi}_{2q+1})$ in (2) there cannot exist a distinct pair $(\underline{\phi}_{2r}, \underline{\phi}_{2r+1})$ from the basis such that

$$(\underline{\phi}_{2r}, \underline{\phi}_{2r+1}) \approx (R^n\underline{\phi}_{2q}, R^n\underline{\phi}_{2q+1}).$$

The only possibility then is that each vector pair must itself be responsible for its rotations, i.e. for each q

$$(\underline{\phi}_{2q}, \underline{\phi}_{2q+1}) \approx (R^n\underline{\phi}_{2q}, R^n\underline{\phi}_{2q+1}) \quad \text{for } n = 0, \ldots, L-1. \quad (3)$$

A model basis which satisfies (3) is said to be *pairwise rotation invariant.*

Our next objective is to investigate conditions under which pairwise rotation invariant model bases exist, and to characterise such bases. Note that as a special case of (3) one has that, for each q

$$\phi_{i,2q}^2 + \phi_{i,2q+1}^2 = (R^n\underline{\phi}_{2q})_i^2 + (R^n\underline{\phi}_{2q+1})_i^2 \quad \text{for } n = 0, \ldots, L-1,$$

from which it follows that, for each q, $(\phi_{i,2q}^2 + \phi_{i,2q+1}^2)$ is constant for $i = 1, \ldots, L$. In fact since $\underline{\phi}_{2q}$ and $\underline{\phi}_{2q+1}$ are normalised

$$\phi_{i,2q}^2 + \phi_{i,2q+1}^2 = \frac{2}{L} \quad \text{for } i = 1, \ldots, L;\ q = 1, \ldots, \frac{1}{2}(L-1).$$

This suggests joining the two elements $\phi_{i,2q}$ and $\phi_{i,2q+1}$ into a single member of $\Re^2$. Let

$$z_{iq} = (\phi_{i,2q}; \phi_{i,2q+1}) \quad \text{for } i = 1, \ldots, L;\ q = 1, \ldots, \frac{1}{2}(L-1). \tag{4}$$

Then for each q the z_{iq}, $i = 1, \ldots, L$ all lie on a circle of radius $\sqrt{2/L}$.

Having established that for each q, the z_{iq} $(i = 1, \ldots, L)$ all lie on the same circle, the full statement of (3) can be used to determine how they are *spaced* on the circle. Note firstly that for any two points z_{iq} and z_{jq} in $\Re^2$ which lie on the same circle, there always exists an $\alpha \in \Re^2$, $|\alpha| = 1$ such that

$$\alpha z_{iq} = z_{jq}.$$

(The definition of the product of two elements in $\Re^2$ is given in the appendix following the proof of Theorem 1.) The next proposition says that if (3) holds then, for each q, there exists a single $\alpha \in \Re^2$, $|\alpha| = 1$ which will transform any z_{iq} to the next one. This means that for each q, the z_{iq} $(i = 1, \ldots, L)$ are equally spaced about the circle.

Proposition 2. A basis Φ, whose elements have been formed into the pairs z_{iq} as in (4) is pairwise rotation invariant iff for each $q = 1, \ldots, \frac{1}{2}(L-1)$ there exists an $\alpha \in \Re^2$, $|\alpha| = 1$ such that

$$\alpha z_{iq} = z_{\{i+1\},q} \quad \text{for } i = 1, \ldots, L \tag{5}$$

where the curly brackets indicate that the quantity within is taken modulo L. •

The z_{iq} can be expressed in terms of their polar representation, i.e.

$$z_{iq} = \sqrt{\frac{2}{L}}\Big(\cos \operatorname{Arg} z_{iq};\ \sin \operatorname{Arg} z_{iq}\Big) \tag{6}$$

where $\operatorname{Arg} z_{iq}$ is the angle which z_{iq} subtends at the origin with the horizontal axis, measured anti-clockwise and chosen such that

$$0 \leq \operatorname{Arg} z_{iq} < 2\pi.$$

Since the length of each z_{iq} is known, z_{iq} is determined completely by the value of its argument. Proposition 2 can be re-stated in terms of the arguments of the z_{iq} .

Lemma 3. A basis Φ is pairwise rotation invariant iff for $q = 1, \ldots, \frac{1}{2}(L-1)$

$$\text{Arg}\, z_{iq} = \{a(i-1)+b\}\frac{2\pi}{L} \quad \text{for } i = 1, \ldots, L \tag{7}$$

for some $a, b \in \Re$. •

The curly brackets (i.e. modulo L) in (7) are redundant in the sense that

$$f(\{a(i-1)+b\}\frac{2\pi}{L}) = f((a(i-1)+b)\frac{2\pi}{L})$$

where f is either cos or sin. Thus without loss of generality we can say that for odd L , a pairwise rotation invariant basis will consist (besides $\sqrt{\frac{1}{L}}1_L$) of pairs of vectors of the form

$$\left[\sqrt{\frac{2}{L}}\cos((a_q i + b_q)\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin((a_q i + b_q)\frac{2\pi}{L})\right]_{i=0,\ldots,L-1} \quad \text{for } q = 1, \ldots, \frac{1}{2}(L-1).$$

The next lemma leads to considerable simplification in that it enables us to discard the terms b_q in the above.

Lemma 4. For all $a, b \in \Re$

$$\left[\cos((ai+b)\frac{2\pi}{L});\ \sin((ai+b)\frac{2\pi}{L})\right]_{i=0,\ldots,L-1} \approx \left[\cos(ai\frac{2\pi}{L});\ \sin(ai\frac{2\pi}{L})\right]_{i=0,\ldots,L-1}.$$

•

We have thus established that, up to $\approx$ −equivalence , a pairwise rotation invariant basis, for odd L , must be of the form

$$\left(\frac{1}{\sqrt{L}}1_L,\ \left[\sqrt{\frac{2}{L}}\cos(a_1 i\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin(a_1 i\frac{2\pi}{L})\right]_i, \cdots \right.$$
$$\left. \cdots, \left[\sqrt{\frac{2}{L}}\cos(a_{\frac{1}{2}(L-1)} i\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin(a_{\frac{1}{2}(L-1)} i\frac{2\pi}{L})\right]_i\right) \tag{8}$$

Now consider choosing the a_q in (8). Note firstly that we can without loss assume that $0 \leq a_q < L$ for each of the a_q. The major consideration in the choice will be that of obtaining an orthogonal basis. The orthogonality conditions which the vectors in (8) must satisfy are:

$$(I) \qquad \sum_{i=0}^{L-1} \cos(a_q i \frac{2\pi}{L}) = 0 \quad \text{for all } a_q$$

$$(II) \qquad \sum_{i=0}^{L-1} \sin(a_q i \frac{2\pi}{L}) = 0 \quad \text{for all } a_q$$

$$(III) \qquad \sum_{i=0}^{L-1} \cos^2(a_q i \frac{2\pi}{L}) = \frac{L}{2} \quad \text{for all } a_q$$

$$(IV) \qquad \sum_{i=0}^{L-1} \sin^2(a_q i \frac{2\pi}{L}) = \frac{L}{2} \quad \text{for all } a_q$$

$$(V) \qquad \sum_{i=0}^{L-1} \cos(a_q i \frac{2\pi}{L}) \cdot \sin(a_r i \frac{2\pi}{L}) = 0 \quad \text{for all } a_q, a_r$$

$$(VI) \qquad \sum_{i=0}^{L-1} \cos(a_q i \frac{2\pi}{L}) \cdot \cos(a_r i \frac{2\pi}{L}) = 0 \quad \text{for all } a_q, a_r; a_q \neq a_r$$

$$(VII) \qquad \sum_{i=0}^{L-1} \sin(a_q i \frac{2\pi}{L}) \cdot \sin(a_r i \frac{2\pi}{L}) = 0 \quad \text{for all } a_q, a_r; a_q \neq a_r$$

Lemma 5. Let $0 \leq a_q < L$. Conditions (I) and (II) are satisfied simultaneously iff $a_q \in \{1, \ldots, L-1\}$. •

It follows that in order to construct an orthogonal pairwise rotation invariant basis we must (in (7)) choose $\frac{1}{2}(L-1)$ unique a_q from $\{1,\ldots,L-1\}$. Note that if $a_q + a_r = L$ then

$$\cos(a_q i\frac{2\pi}{L}) = \cos(a_r i\frac{2\pi}{L})$$

and

$$\sin(a_q i\frac{2\pi}{L}) = -\sin(a_r i\frac{2\pi}{L})$$

so that the vector pairs

$$\left[\cos(a_q i\frac{2\pi}{L});\ \sin(a_q i\frac{2\pi}{L})\right]_{i=0,\ldots,L-1}$$

and

$$\left[\cos(a_r i\frac{2\pi}{L});\ \sin(a_r i\frac{2\pi}{L})\right]_{i=0,\ldots,L-1}$$

are non-distinct. Thus, without loss of generality, we can choose the first $\frac{1}{2}(L-1)$ elements from $\{1,\ldots,L-1\}$ so that

$$a_q = q \quad \text{for } q = 1,\ldots,\frac{1}{2}(L-1).$$

It can be shown, see Bloomfield (1976), that this choice does lead to the basis in (7) being orthogonal.

We have thus proved:

Theorem 6. There is a $\approx$ −unique pairwise rotation invariant model basis for each odd L, namely

$$\left(\frac{1}{\sqrt{L}}1_L,\ \left[\sqrt{\frac{2}{L}}\cos(1i\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin(1i\frac{2\pi}{L})\right]_i,\ \left[\sqrt{\frac{2}{L}}\cos(2i\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin(2i\frac{2\pi}{L})\right]_i,\ldots\right.$$

$$\left.\ldots,\left[\sqrt{\frac{2}{L}}\cos(\frac{1}{2}(L-1)i\frac{2\pi}{L});\ \sqrt{\frac{2}{L}}\sin(\frac{1}{2}(L-1)i\frac{2\pi}{L})\right]_i\right). \qquad (9)$$

•

We turn now to even L and deal with that spare last vector in the basis. Orthonormality and pairwise rotation invariance are obtained if for the paired vectors we use the sin-cos pairs as above, while the last vector is taken to be (the normalised form of)

$$\left[\cos((\frac{1}{2}L)i\frac{2\pi}{L})\right]_i = \left[\cos(i\pi)\right]_i = \underline{\Lambda}_L.$$

This is in fact, the only vector, up to $\sim$ −equivalence , with the required properties. We can thus state:

Theorem 7. Up to $\approx$ −equivalence of the paired vectors and $\sim$ −equivalence of the non-paired vector, there is a unique pairwise rotation invariant model basis for each even L , namely

$$\Big(\frac{1}{\sqrt{L}}\underline{1}_L, \left[\sqrt{\frac{2}{L}}\cos(1i\frac{2\pi}{L}); \sqrt{\frac{2}{L}}\sin(1i\frac{2\pi}{L})\right]_i, \ldots$$
$$\ldots, \left[\sqrt{\frac{2}{L}}\cos(\frac{1}{2}(L-2)i\frac{2\pi}{L}); \sqrt{\frac{2}{L}}\sin(\frac{1}{2}(L-2)i\frac{2\pi}{L})\right]_i, \frac{1}{\sqrt{L}}\underline{\Lambda}_L\Big). \qquad (10)$$

•

Bases of the form (9) and (10) are the well-known Fourier bases. We have established then that (i) a pairwise rotation invariant model basis exists for every dimension, (ii) that for each dimension there is essentially only one such basis; and (iii) this basis can be taken to be the Fourier basis.

Notation. 1. The Fourier basis for a variable with L categories will be denoted by F_L .

2. It is possible to represent the pairwise modelling procedure for odd and even L using the same expressions. Note that for even L

$$\left[\sin(\frac{L}{2}i\frac{2\pi}{L})\right]_{i=0,\ldots,L-1} = \underline{0}\,.$$

This vector can then be paired up with the "spare" last vector which occurs in Fourier bases for even L , to get

$$\left[\cos(\frac{L}{2}i\frac{2\pi}{L}); \sin(\frac{L}{2}i\frac{2\pi}{L})\right]_{i=0,\ldots,L-1}.$$

Including this extra vector has no effect on the modelling procedure, (the corresponding parameter and its contribution to the estimated expected discrepancy are both zero), but it does allow us to write all models, for odd or even L, in terms of vector pairs. (See the summary below.)

3. If we define

$$[L/2] = \begin{cases} L/2 & \text{if } L \text{ is even} \\ (L-1)/2 & \text{if } L \text{ is odd} \end{cases}$$

then all saturated models will have $[L/2]$ parameter pairs.

4. Since the i subscript in the cos-sin pair runs from 0 to $L-1$, we will label the cells in the classification from 0 to $L-1$ and adopt the convention that the i subscript will run from 0 to $L-1$, (in particular in $\pi_i, M_i(\theta)$ and P_i).

Summary. If we have a cyclical variable with any number of categories, L, then we can use the pairwise modelling procedure in conjunction with the appropriate Fourier basis, so that the fitted model obtained is

$$M_i(\hat{\theta}) = \frac{1}{L} + \sum_{q \in Q} \left[\hat{\theta}_{2q} \cos(iq\frac{2\pi}{L}) + \hat{\theta}_{2q+1} \sin(iq\frac{2\pi}{L}) \right] \quad \text{for } i = 0, \ldots, L-1 \qquad (11)$$

where

$$\hat{\theta}_{2q} = \sum_{i=0}^{L-1} \cos(iq\frac{2\pi}{L}) P_i; \quad \hat{\theta}_{2q+1} = \sum_{i=0}^{L-1} \sin(iq\frac{2\pi}{L}) P_i$$

$$Q \subseteq \{1, \ldots, [\frac{L}{2}]\}$$

$$q \in Q \quad \text{iff} [\hat{\theta}_{2q}^2 - 2\widehat{\text{var}}\,\hat{\theta}_{2q}] + [\hat{\theta}_{2q+1}^2 - 2\widehat{\text{var}}\,\hat{\theta}_{2q+1}] \leq 0.$$

Henceforth "pairwise modelling procedure" will refer to the modelling procedure which produces the model (11).

6.4. APPLICATIONS AND EXTENSIONS

In this section we give some examples of application of the theory developed in the previous section. Specifically we

(1) provide an alternative parameterisation for models using Fourier bases,

(2) fit a number of models for univariate cyclical variables using the pairwise modelling procedure,

(3) give the extension to multiway cross-classifications involving some cyclical variables, and

(4) give examples of the extension.

The amplitude-phase representation. The individual parameter values/estimates obtained when using the pairwise modelling procedure vary with the particular rotation of the cells that is used. However it is possible to reparameterise the model in such a way that the parameter pair $(\theta_{2q}, \theta_{2q+1})$ are transformed to (A_q, ρ_q) where A_q is rotation invariant and ρ_q , although not rotation invariant, varies by a fixed amount for each rotation.

The reparameterisation is achieved by writing

$$\theta_{2q}\cos(iq\frac{2\pi}{L}) + \theta_{2q+1}\sin(iq\frac{2\pi}{L}) = A_q\cos(iq\frac{2\pi}{L} - \rho_q) \qquad (1)$$

for some (initially unknown) A_q and ρ_q . A_q and ρ_q are found by expanding the left side of (1) as

$$A_q[\cos(iq\frac{2\pi}{L})\cos\rho_q + \sin(iq\frac{2\pi}{L})\sin\rho_q]$$

from which it follows that

$$\theta_{2q} = A_q\cos\rho_q$$

$$\theta_{2q+1} = A_q\sin\rho_q.$$

Solving for A_q and ρ_q in terms of θ_{2q} and θ_{2q+1} yields

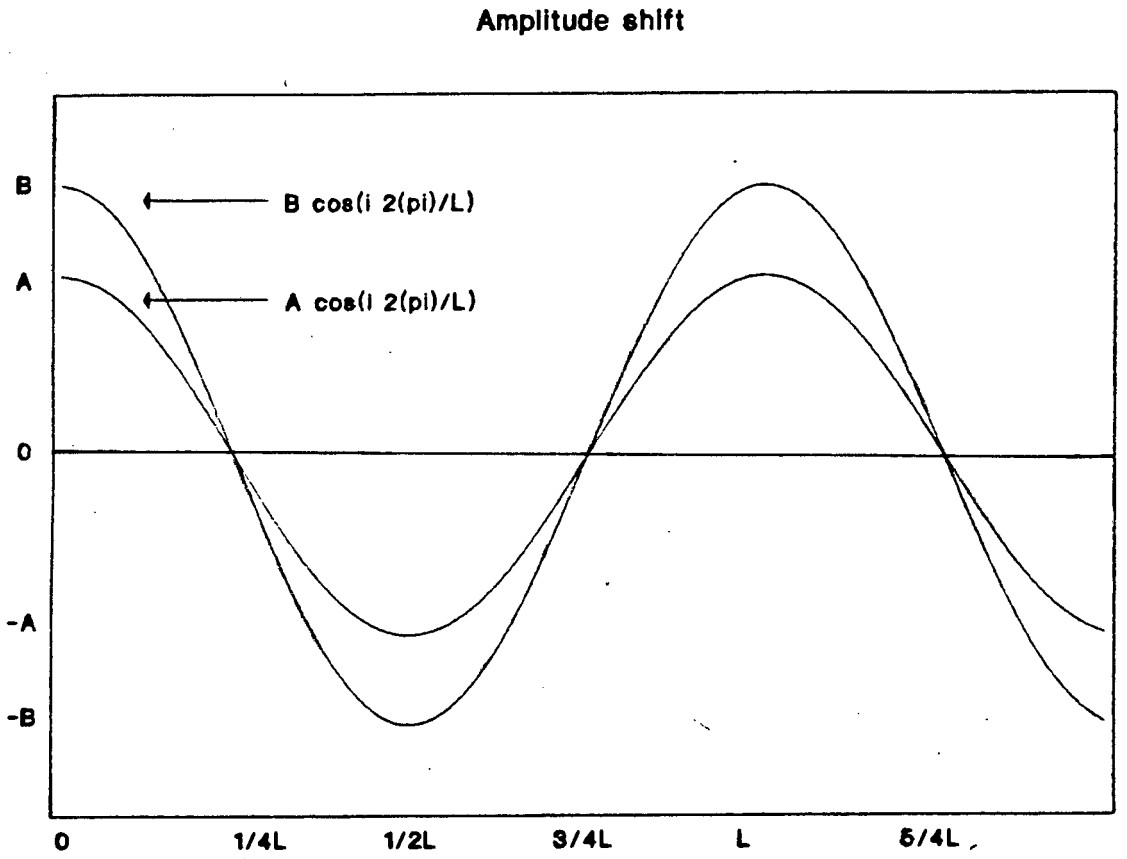
$$A_q = (\theta_{2q}^2 + \theta_{2q+1}^2)^{\frac{1}{2}}$$

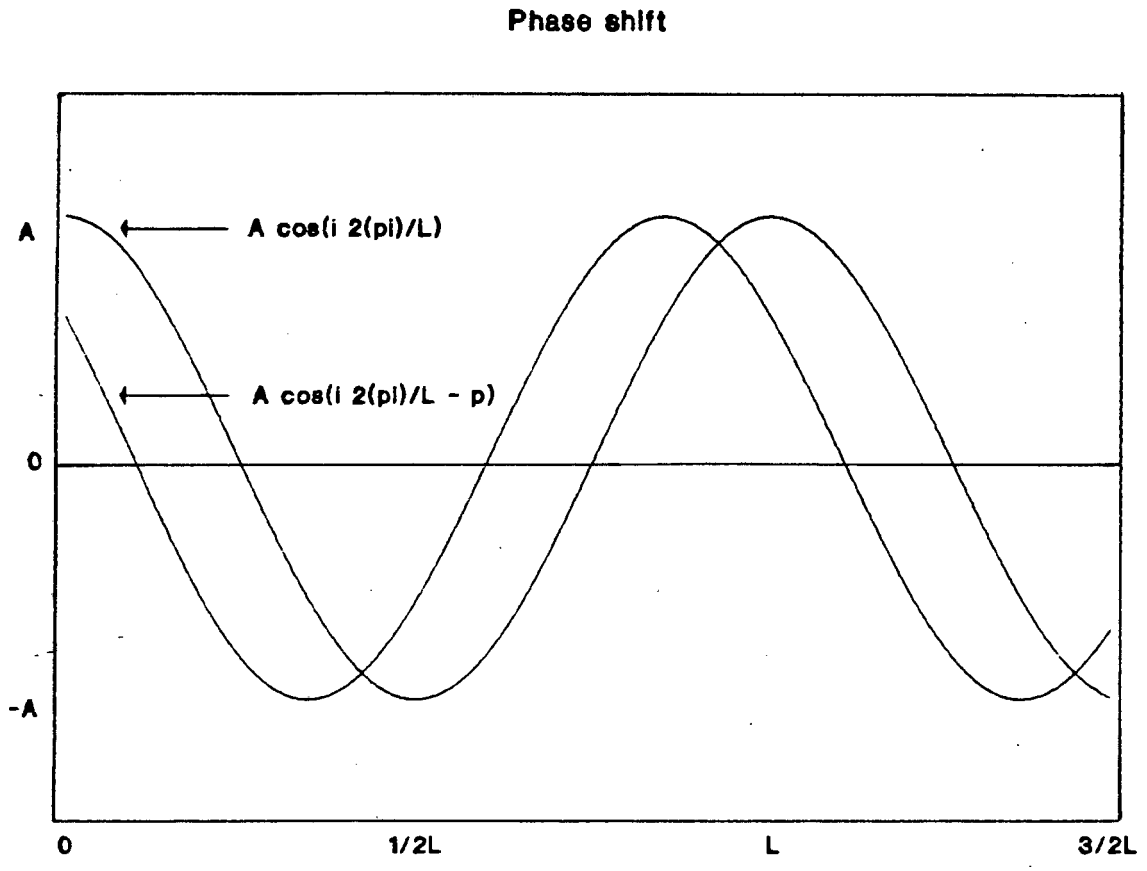
$$\tan \rho_q = \theta_{2q+1}/\theta_{2q}\,.$$

In order that the ρ_q lie in the interval $[0, 2\pi)$ we use the following convention to compute the ρ_q .

$$\rho_q = \begin{cases} \arctan\,(\theta_{2q+1})/(\theta_{2q}) & \text{if } \theta_{2q} > 0,\ \theta_{2q+1} \geq 0 \\ \arctan\,(\theta_{2q+1})/(\theta_{2q}) + 2\pi & \text{if } \theta_{2q} > 0,\ \theta_{2q+1} < 0 \\ \arctan\,(\theta_{2q+1})/(\theta_{2q}) + \pi & \text{if } \theta_{2q} < 0 \\ (1/2)\pi & \text{if } \theta_{2q} = 0,\ \theta_{2q+1} \geq 0 \\ (3/2)\pi & \text{if } \theta_{2q} = 0,\ \theta_{2q+1} < 0. \end{cases}$$

Sketches of $A_q \cos(iq\frac{2\pi}{L} - \rho_q)$ viewed as continuous functions in i are given for a few values of q , and are used in explaining the interpretation placed on A_q and ρ_q .





A_q represents the maximum height which the graph of $A_q \cos(iq\frac{2\pi}{L} - \rho_q)$ achieves and is called the *amplitude.* A_q is invariant with respect to rotations of the cells. (This is easily shown and is a direct consequence of the basis being rotation invariant.)

ρ_q determines the value that the graph has when $i = 0$ and is called the *phase.* The phase is, of course not invariant with respect to rotations. However using the above definition ρ_q changes by $2\pi/L$ radians with every rotation.

Note that the wave associated with $A_q \cos(iq\frac{2\pi}{L} - \rho_q)$ completes exactly q cycles as i runs from 0 to L, and consequently q is called the (Fourier) *frequency*.

A_q and ρ_q will be estimated by their maximum likelihood estimates, i.e. by replacing θ_{2q} and θ_{2q+1} in the expressions defining A_q and ρ_q by their maximum likelihood estimators $\hat{\theta}_{2q}$ and $\hat{\theta}_{2q+1}$.

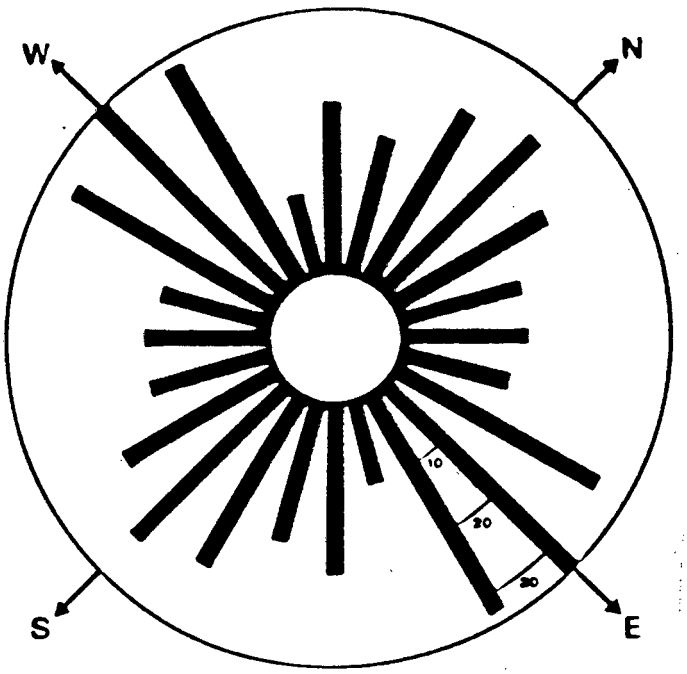
Parameter and criterion tables. Since the individual parameter estimates, as well as their standard deviations and contributions to the expected discrepancy, are not rotation invariant, they are not given in the parameter and criterion tables. The two sets of quantities that are rotation invariant and which are given instead, are

(i) the amplitudes, A_q, for $q = 1, \ldots, [\frac{L}{2}]$

(ii) the joint contribution of each cos-sin pair to the expected discrepancy.

That the joint contributions which are just sums of the two individual contributions, should be rotation invariant is one of the requirements for the pairwise modelling procedure to be rotation invariant.

Example. A circular histogram (Batschelet 1981) gives counts of the orientation of resting flies of the species *Calliphora erythrocephala* in each of the 24 direction segments.

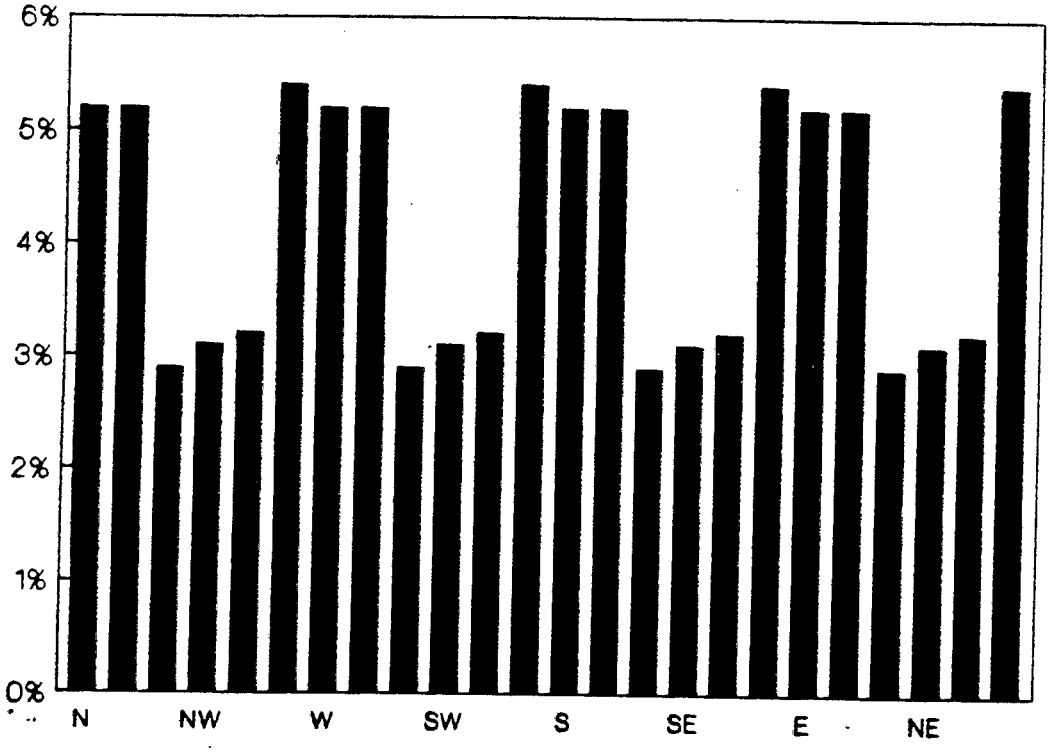


In modelling these observations we clearly want to use a rotation invariant procedure. Since there are 24 cells we must use the pairwise modelling procedure in conjunction with the Fourier basis, F_{24} .

The amplitude and criterion table

	Fourier frequency											
	1	2	3	4	5	6	7	8	9	10	11	12
amplitude ($\times 10^3$)	2.3	16.2	0.5	51.2	1.0	17.1	0.9	16.0	0.5	14.5	0.6	19.4
contribution ($\times 10^3$)	0.3	0.0	0.3	-2.3	0.3	0.0	0.3	0.0	0.3	0.1	0.3	-0.2

There are only two negative contributions. The first of these comes from the fourth cos-sin pair. Including the corresponding pair of parameters in the model will mean that the magnitudes of the fitted probabilities will exhibit a wave-like structure, where the wave is repeated four times among the twenty four points, with a cycle length of six. In fact the magnitudes of the fitted probabilities follow the pattern:



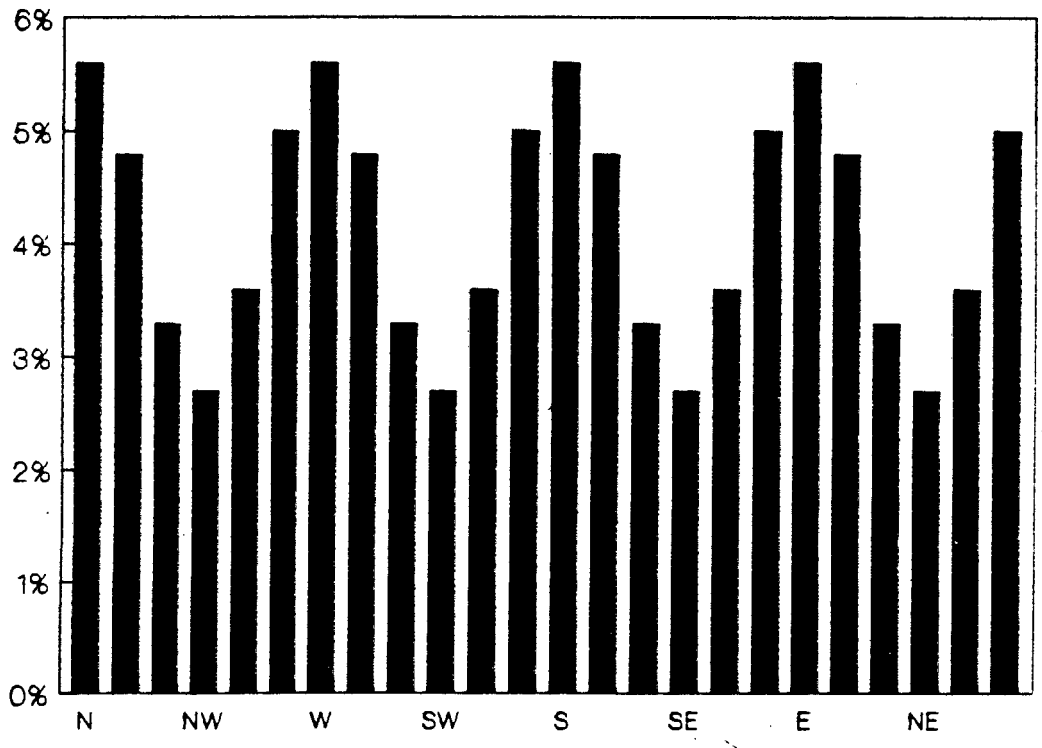
The remaining negative contribution is the last in the list. This corresponds not to a pair, but to the single vector

$$\left[\cos\left(\frac{12}{2}i\frac{2\pi}{12}\right)\right]_{i=0,\ldots,L-1} = \underline{\Lambda}_{24}$$

which nevertheless gives a wave whose frequency is twelve with a cycle length of two. Including the associated parameter into the existing model means that

$$\frac{1}{\sqrt{24}}\hat{\theta}_{24} = \frac{1}{24}\sum_{i=0}^{23}(-1)^i P_i$$

is added to the first cell, subtracted from the next, added to the next, and so on. Note that in particular this means that every sixth fitted probability gets altered by the same amount, so that the fitted probabilities will still be repeated after every six. The pattern which the fitted probabilities now follow is:



Example. Plackett (1974) gives a data set wherein all the cases of acute lymphatic leukaemia reported to the British National Cancer Registration Scheme during 1946-60 were classified by month of clinical onset.

Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
40	34	30	34	39	58	51	55	36	48	33	38

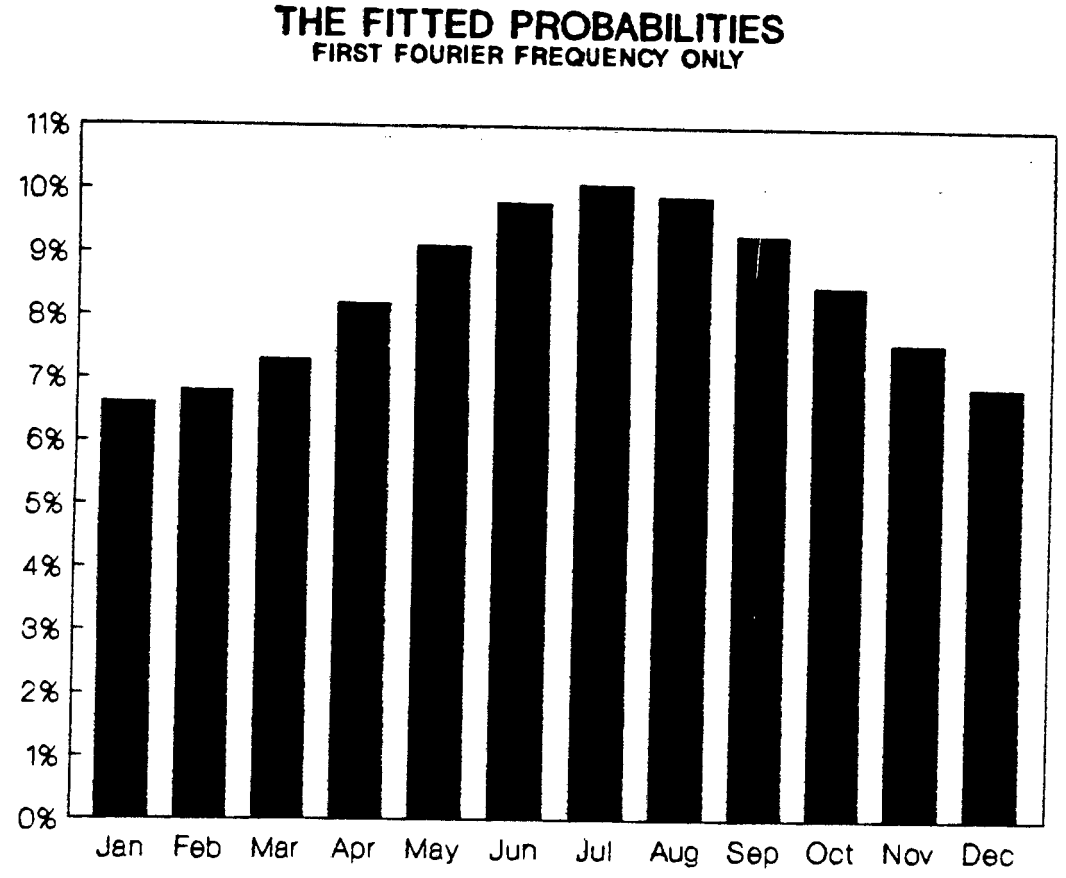
The operating model is not multinomial as the total sample was not fixed as part of the experimental design, but instead represents the realisation of a random variable. The appropriate operating model is thus Poisson. However it is well known that the distribution of Poisson cell counts conditional on the sum of the cell counts is multinomial (see, for example, Plackett (1977, p.4)). We will thus model the counts conditional on their sum.

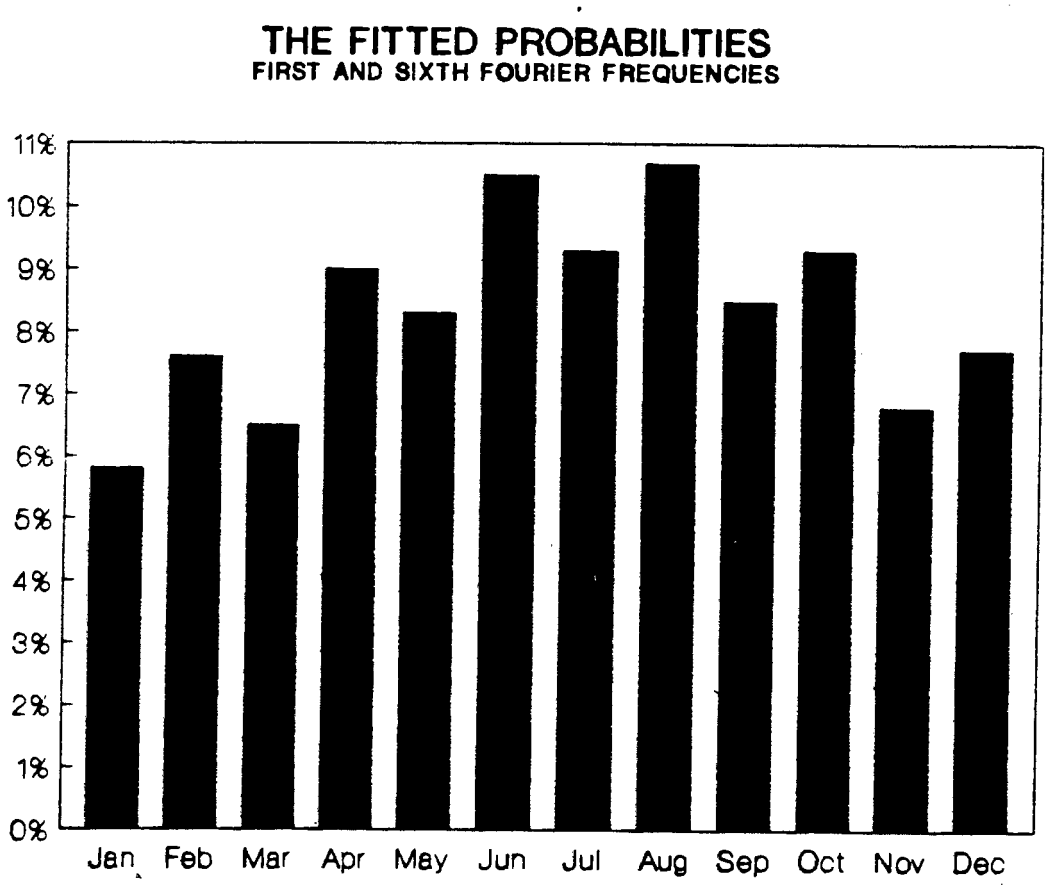
Since there are 12 cells we use the pairwise modelling procedure in conjunction with F_{12} .

The amplitude and criterion table

	Fourier frequency					
	1	2	3	4	5	6
amplitude	42.5	20.3	2.6	17.4	15.4	27.4
contribution	-1.2	0.2	0.7	0.4	0.4	-0.4

There are two negative contributions. Including the parameter pair associated with the first of these means that the twelve fitted probabilities will resemble a wave with a cycle length of twelve, which indicates that the number of cases of leukemia reported follows an annual cycle. Including the parameter associated with the highest frequency wave has the usual "up-down up-down" effect.


THE FITTED PROBABILITIES
FIRST FOURIER FREQUENCY ONLY
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec


THE FITTED PROBABILITIES
FIRST AND SIXTH FOURIER FREQUENCIES
11%
10%
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

More than one variable. The situation regarding multiway cross-classifications where some or all of the variables are cyclical, is easily dealt with.

Consider two random variables X and Y; with R and C categories respectively and with their respective model bases Ψ and Ω. Suppose that Y is a cyclical variable. The obvious generalisation of the univariate pairwise modelling procedure is that which produces fitted models of the form

$$M_{ij}(\hat{\theta}) = \sum_{r \in A} \psi_{ir} \frac{1}{\sqrt{C}} \hat{\theta}_{r1} + \sum_{(r,c) \in B} \psi_{ir} \{\omega_{j,2c} \hat{\theta}_{r,2c} + \omega_{j,2c+1} \hat{\theta}_{r,2c+1}\}$$

where

(i) for $r = 1, \ldots, R;\ r \in A$ iff

$$(\hat{\theta}_{r1})^2 - 2 \widehat{\text{var}}\, \hat{\theta}_{r1} < 0$$

(ii) for $r = 1, \ldots, R;\ c = 1, \ldots, [L/2];\ (r,c) \in B$ iff

$$\{\hat{\theta}_{r,2c} - 2 \widehat{\text{var}}\, \hat{\theta}_{r,2c}\} + \{\hat{\theta}_{r,2c+1} - 2 \widehat{\text{var}}\, \hat{\theta}_{r,2c+1}\} < 0$$

(iii) $\hat{\theta}_{r,L+1}$ and $\text{var}\, \hat{\theta}_{r,L+1}$ are defined to be zero.

It is not difficult to show that if Ω is a pairwise rotation invariant basis then this modelling procedure is invariant with respect to rotations of the Y categories.

The above considerations generalise to cross-classifications involving more than two variables. No matter how many variables appear in a cross-classification, for each of the variables which is cyclical one has simply to use a pairwise rotation invariant model basis and the pairwise selection criterion. This leads to a procedure which is invariant to rotations of the categories of the cyclical variables.

Example. Linhart and Zucchini (1986a) analyse a data set involving the rate of arrival of storms at different times of the year. The table below gives, for each week of the year, the number of times that at least one storm arrived at the Botanic Gardens, Durban for the period 1.6.1932–31.12.1979. The definition used for a "storm" is that at least 30 mm of rain fell within a twenty-four period.

Week	Begin	Storm Yes	Storm No	Week	Begin	Storm Yes	Storm No
1	1 Jan	6	41	27	2 Jul	4	44
2	8 Jan	8	39	28	9 Jul	0	48
3	15 Jan	7	40	29	16 Jul	2	46
4	22 Jan	6	41	30	23 Jul	0	48
5	29 Jan	9	38	31	30 Jul	3	45
6	5 Feb	15	32	32	6 Aug	1	47
7	12 Feb	6	41	33	13 Aug	1	47
8	19 Feb	12	35	34	20 Aug	5	43
9*	26 Feb	16	31	35	27 Aug	4	44
10	5 Mar	7	40	36	3 Sep	3	45
11	12 Mar	9	38	37	10 Sep	6	42
12	19 Mar	6	41	38	17 Sep	1	47
13	26 Mar	8	39	39	24 Sep	8	40
14	2 Apr	2	45	40†	1 Oct	3	45
15	9 Apr	7	40	41	9 Oct	4	44
16	16 Apr	4	43	42	16 Oct	6	42
17	23 Apr	4	43	43	23 Oct	9	39
18	30 Apr	3	44	44	30 Oct	5	43
19	7 May	3	44	45	6 Nov	8	40
20	14 May	10	37	46	13 Nov	6	42
21	21 May	3	44	47	20 Nov	5	43
22	28 May	3	44	48	27 Nov	7	41
23	4 Jun	0	48	49	4 Dec	5	43
24	11 Jun	5	43	50	11 Dec	8	40
25	18 Jun	1	47	51	18 Dec	5	43
26	25 Jun	2	46	52	25 Dec	4	44

* Eight days on leap year † Eight days

In selecting a model for this data set, a modelling procedure which is invariant to the choice of which week is labelled as the first, is required. Consequently the basis used in connection with the variable representing weeks is the Fourier basis, F_{52} , and the pairwise selection criterion is employed.

For the presence/absence variable H_2 is used. (As is expected the modelling procedure is invariant to swopping the order of the "yes" and "no" categories. This because H_2 is a rotation invariant model basis.)

Note that the number of weeks on which observations were taken is fixed. Consequently the operating model is product-multinomial with weeks the explanatory variable whose category totals are fixed, so that there are only 52 linearly independent parameters.

The amplitude and criterion table

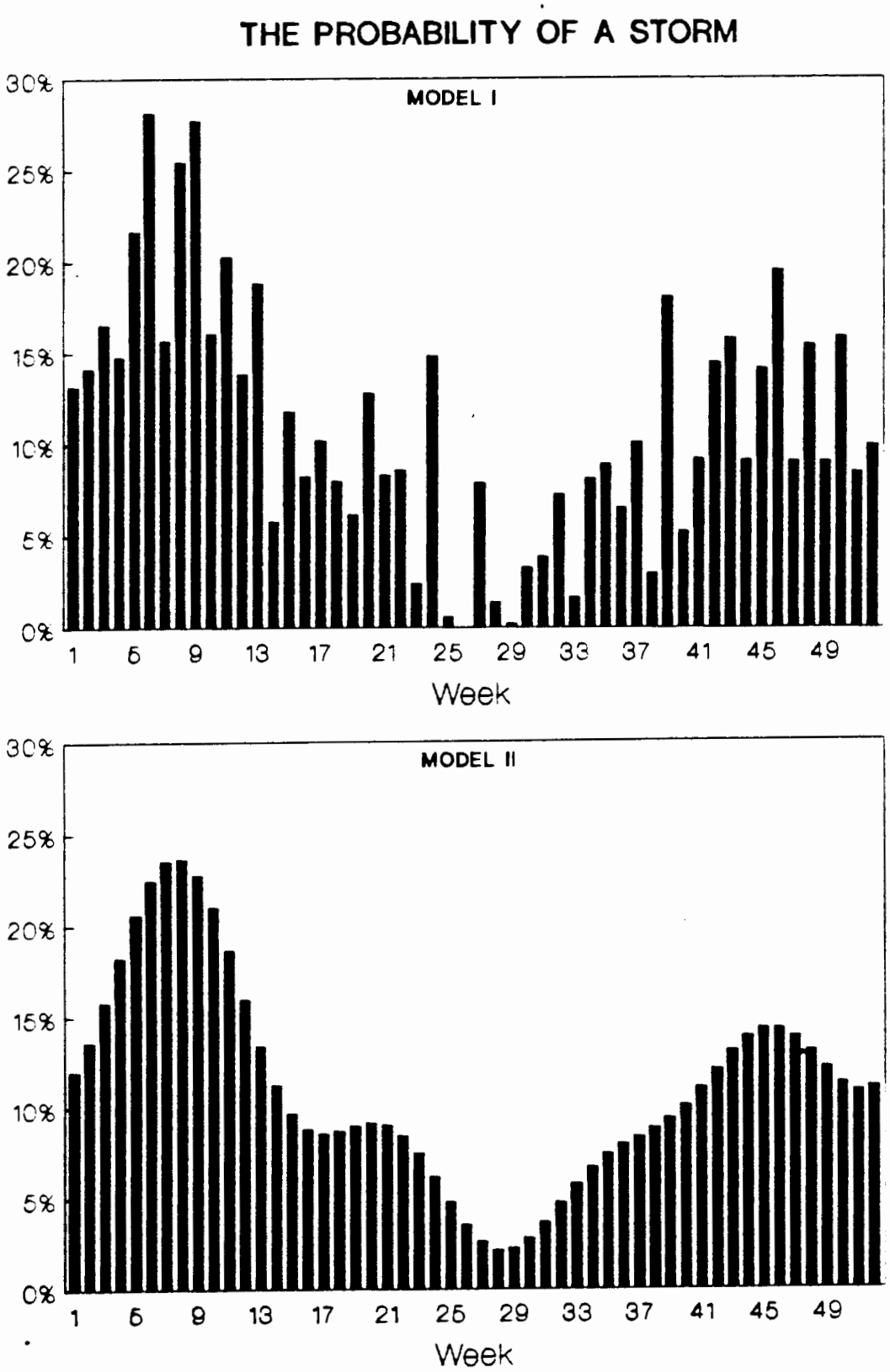
Fourier frequency												
1	2	3	4	5	6	7	8	9	10	11	12	13
489.2	170.9	181.5	157.2	49.2	73.8	49.2	29.4	91.7	37.6	89.0	5.2	19.1
-223.2	-13.1	-16.8	-8.6	13.7	10.7	13.7	15.3	7.7	14.7	8.2	16.1	15.8

Fourier frequency												
14	15	16	17	18	19	20	21	22	23	24	25	26
205.1	115.4	44.6	148.1	65.8	101.5	129.1	10.1	179.2	117.3	174.4	57.0	45.1
-25.9	2.8	14.2	-5.8	11.8	5.9	-0.5	16.1	-15.9	2.4	-14.3	12.9	6.0

It is seen that the first four parameter pairs, corresponding to the low Fourier frequencies, are marked for inclusion. After these four there are a number of parameter pairs whose contributions are also negative. When so many parameters are involved, it is quite likely that at least some of the parameter pairs marked for inclusion do not really decrease the actual discrepancy between fitted and operating model, although through sampling variation their contributions to the estimated expected discrepancy are negative. Furthermore one expects the rate of arrival of storms to vary smoothly with time, and hence one should be wary of including the

high frequency terms.

Two models were fitted to the data: Model I, where only the constant term and the first four Fourier frequency parameter pairs were included and Model II where the constant term and all the parameter pairs whose contributions to the estimated expected discrepancy were negative were included.



Example. Consider the wind data again. This time we introduce the hour of the day, and model the conditional probabilities with which the wind blows in each of the sixteen direction segments, where the conditioning event is the hour. The table of counts is taken from Iloni (1986) and involves 43 128 observations collected over the time period 1 January 1979 to 31 October 1984.

December?

	hour 1	hour 2	hour 3	hour 4	hour 5	hour 6	hour 7	hour 8	hour 9	hour 10	hour 11	hour 12
N	95	90	97	92	71	67	74	78	81	106	113	113
	64	43	39	41	40	48	42	53	67	77	88	112
NW	28	32	31	31	35	29	26	38	62	89	135	191
	32	39	33	31	30	40	44	35	59	91	150	217
W	34	36	34	38	24	34	34	40	73	132	189	198
	29	44	36	37	36	31	35	62	97	146	156	202
SW	71	67	54	48	62	46	57	79	126	142	165	170
	86	85	102	106	93	93	96	111	103	84	85	81
S	123	121	125	124	138	134	134	117	99	89	89	71
	168	152	134	153	164	157	156	129	119	90	67	55
SE	102	128	143	122	109	160	154	129	76	53	33	27
	170	183	196	225	243	256	234	247	200	133	59	28
E	112	116	132	153	159	144	162	148	143	121	102	63
	118	130	143	133	144	132	136	102	88	74	88	57
NE	260	255	255	232	220	215	220	192	167	137	80	73
	303	275	240	230	227	210	192	236	236	232	197	139

	hour 13	hour 14	hour 15	hour 16	hour 17	hour 18	hour 19	hour 20	hour 21	hour 22	hour 23	hour 24
N	138	118	158	207	257	303	242	190	149	146	142	117
	147	157	168	161	158	149	110	111	96	83	55	53
NW	230	288	311	320	270	232	180	102	61	39	45	33
	267	264	250	234	192	143	98	71	43	42	25	37
W	212	227	220	177	154	116	79	71	49	39	41	36
	190	164	157	167	160	131	110	64	64	51	44	32
SW	187	198	210	207	203	187	169	121	86	62	61	61
	64	69	61	57	69	106	113	132	126	96	89	92
S	76	58	41	41	40	49	73	87	102	117	117	112
	29	25	19	25	22	19	30	71	101	117	141	161
SE	22	14	19	18	20	29	26	46	71	95	93	111
	21	14	17	15	14	20	26	53	76	91	123	134
E	47	29	25	23	18	18	27	55	72	93	101	112
	50	45	32	31	34	31	48	66	82	120	135	137
NE	38	37	32	23	31	52	96	155	200	217	244	256
	108	89	76	90	155	220	366	402	419	389	341	312

Both of the variables, wind direction and hour of day, are cyclical. For the variable associated with the hour of day the Fourier basis, F_{24} , is used and the vectors are considered pairwise. For the variable associated with the direction categories a rotation invariant model basis of dimension 16 is used and the vectors are considered singly. In order to construct a rotation invariant model basis of dimension 16, one needs, besides $\underline{1}_{16}$, rotation groups of cardinalities 1, 2, 4 and 8. For the first three rotation groups it is proposed that we use essentially the same generators as were used for the $L = 8$ cases, i.e. the generators

$$\underline{\Lambda}_{16};\ \underline{\Lambda}_8 \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \underline{\Lambda}_4 \otimes \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

The generator of cardinality 8 that was used, is

$$\underline{\Lambda}_2 \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note that the number of observations taken at each of the hours was fixed by design and is not something which should be modelled. The operating model is taken to be product-multinomial, with hour the explanatory variable and direction the response variable. The saturated model contains $(16 - 1) \times 24$ free parameters.

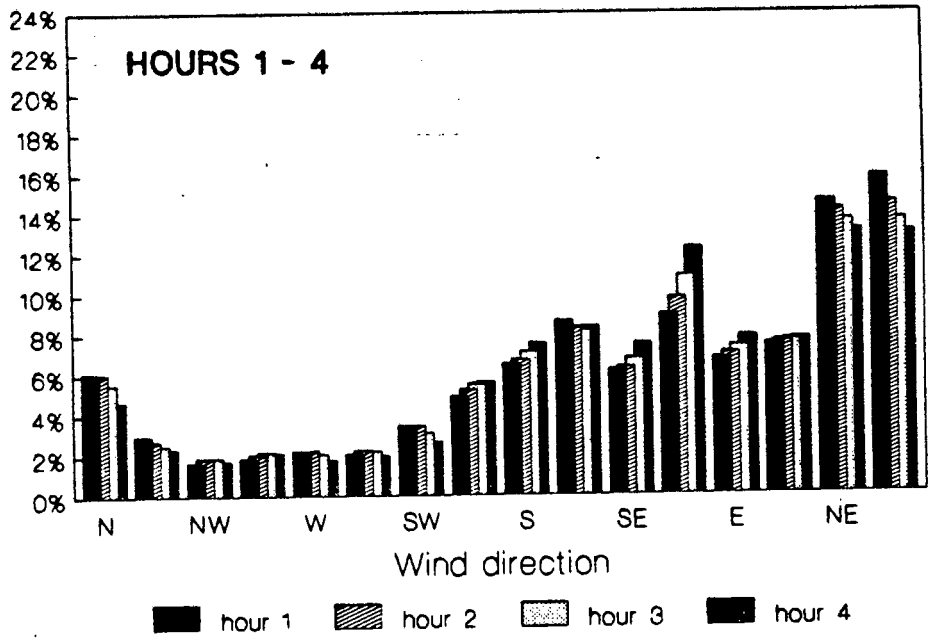
The amplitude table

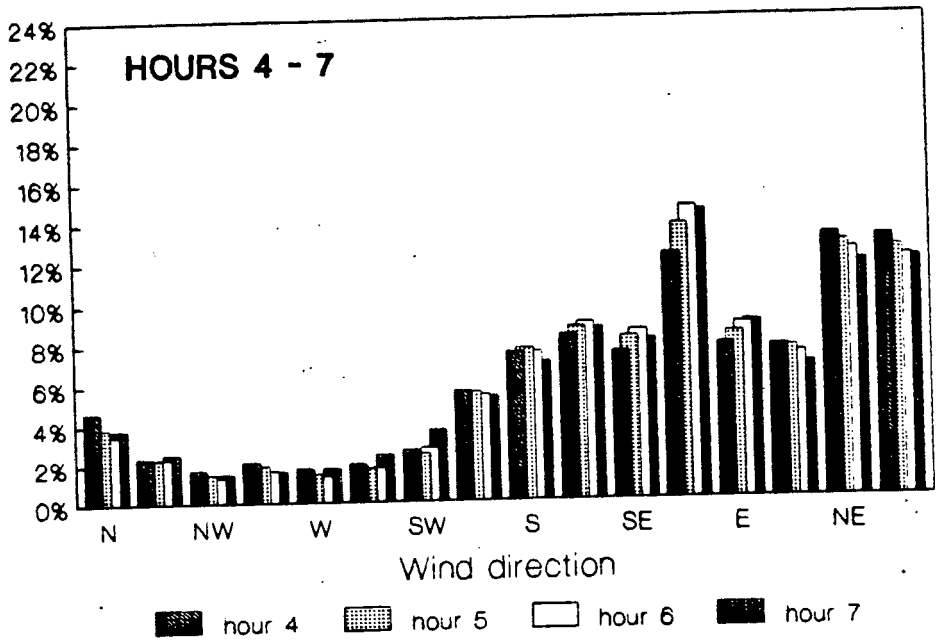
	Fourier frequency											
0	1	2	3	4	5	6	7	8	9	10	11	12
59.6	243.7	49.9	31.8	18.7	5.2	11.1	4.8	7.2	5.7	4.4	3.3	5.8
2.9	256.2	38.4	26.5	2.6	7.7	10.4	4.6	7.6	2.8	5.5	8.4	3.7
11.3	381.1	93.3	29.0	17.4	19.4	4.9	13.3	7.1	2.9	7.0	4.0	6.1
215.8	263.4	123.8	14.3	18.3	15.8	7.9	9.0	9.0	4.6	2.7	13.0	3.5
46.9	71.8	63.6	33.9	14.6	14.3	3.3	2.6	4.6	5.0	0.7	4.3	2.9
13.2	35.1	18.2	13.9	7.3	3.0	8.7	6.1	3.3	5.6	7.6	9.0	2.6
107.6	35.3	22.8	12.0	6.1	6.9	4.5	2.0	0.4	4.8	4.3	9.0	1.3
180.1	186.0	52.3	33.2	6.9	9.3	4.0	5.4	10.0	6.9	8.5	6.0	3.1
34.8	86.0	60.3	10.5	16.2	17.3	8.8	3.9	4.2	6.7	4.1	2.3	6.2
108.4	90.5	89.9	54.1	5.2	4.6	12.1	2.0	4.7	8.4	7.8	5.5	3.9
117.5	116.3	57.1	28.9	6.9	7.9	6.4	3.7	1.6	4.1	2.3	5.4	4.0
108.5	148.0	94.9	24.5	17.8	17.2	6.2	10.7	2.1	4.4	2.7	2.1	0.4
48.5	12.7	27.6	22.1	20.6	4.0	1.8	9.1	6.8	2.9	6.6	6.8	0.9
171.2	63.7	57.3	18.3	27.1	11.1	3.3	3.7	2.3	5.6	4.6	6.7	1.3
25.2	71.2	49.1	20.6	11.3	4.2	18.0	4.6	2.2	3.4	0.6	4.3	1.1

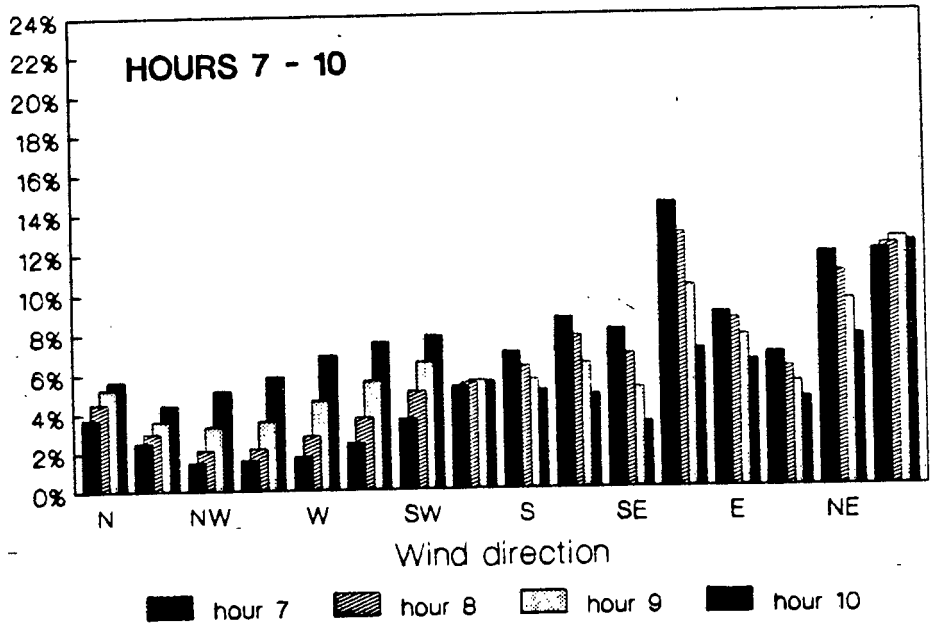
The criterion table

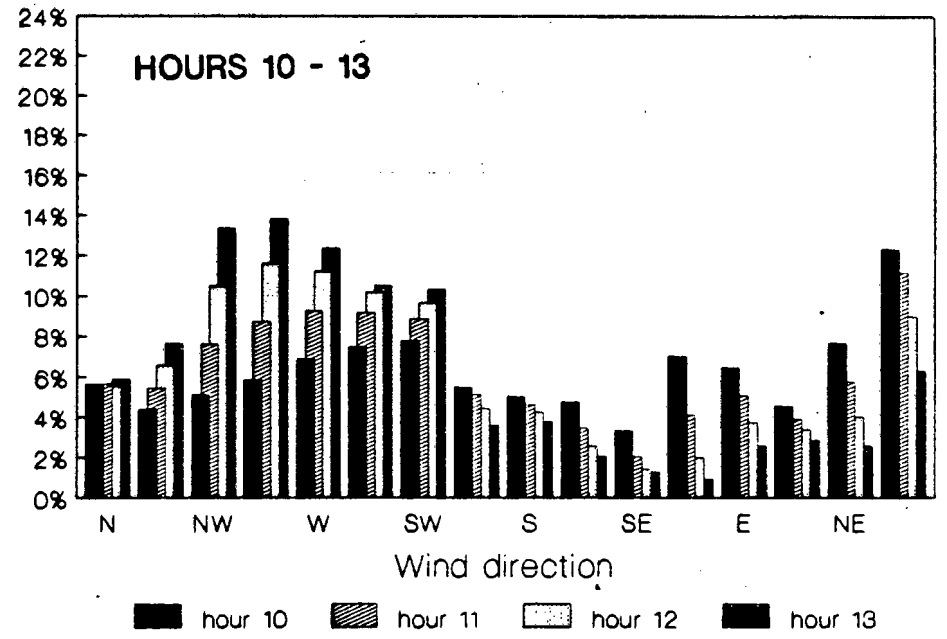
Fourier frequency												
0	1	2	3	4	5	6	7	8	9	10	11	12
-3.5	-59.3	-2.4	-0.9	-0.2	-0.0	-0.0	0.1	0.1	0.1	0.1	0.1	0.0
0.0	-65.5	-1.4	-0.6	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0
-0.1	-145.1	-8.6	-0.7	-0.2	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.0
-46.5	-69.2	-15.2	-0.0	-0.2	-0.2	0.1	0.1	0.1	0.1	0.2	0.0	0.1
-2.1	-5.0	-3.9	-1.0	-0.2	-0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-0.1	-1.1	-0.2	-0.1	0.1	-0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.1
-11.5	-1.1	-0.4	-0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-32.4	-34.3	-2.6	-0.9	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-1.1	-7.3	-3.5	0.0	-0.1	-0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.0
-10.6	-8.0	-7.9	-2.8	-0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1
-13.7	-13.4	-3.1	-0.7	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-11.7	-21.8	-8.9	-0.5	-0.2	-0.2	0.1	0.0	0.1	0.1	0.1	0.1	0.1
-2.3	-0.0	-0.6	-0.4	-0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-29.2	-3.9	-3.1	-0.2	-0.6	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
-0.6	-4.9	-2.3	-0.3	0.0	0.1	-0.2	0.1	0.1	0.1	0.1	0.1	0.1

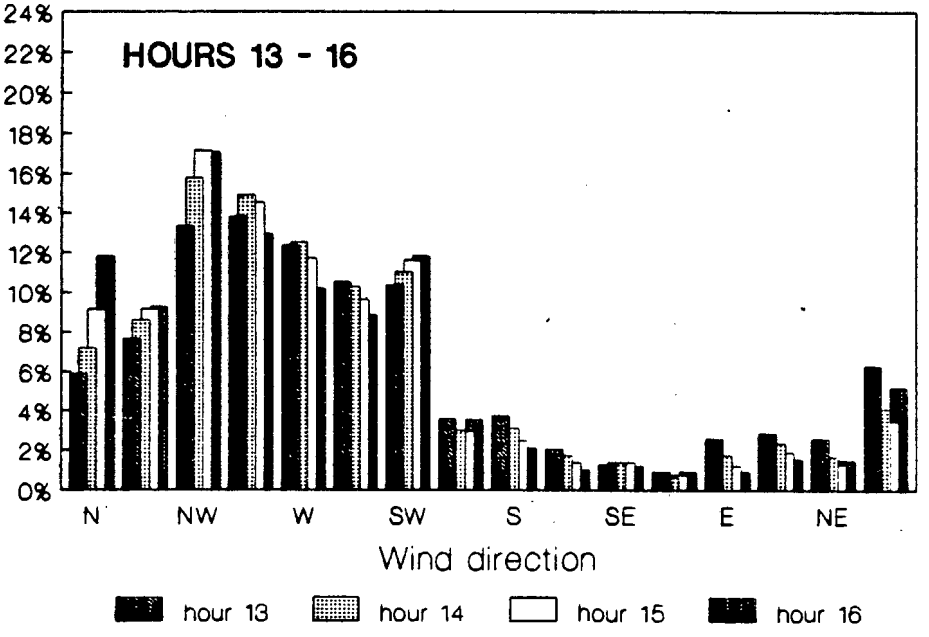
From the criterion table one notes immediately that the contributions corresponding to the high order frequency Fourier vector pairs are small in absolute value and are generally non-negative. In fact it might be advisable to make a blanket exclusion rule, excluding from the model all parameter (-pairs) associated with Fourier frequencies of four or more. The fitted conditional probabilities for this model are shown.

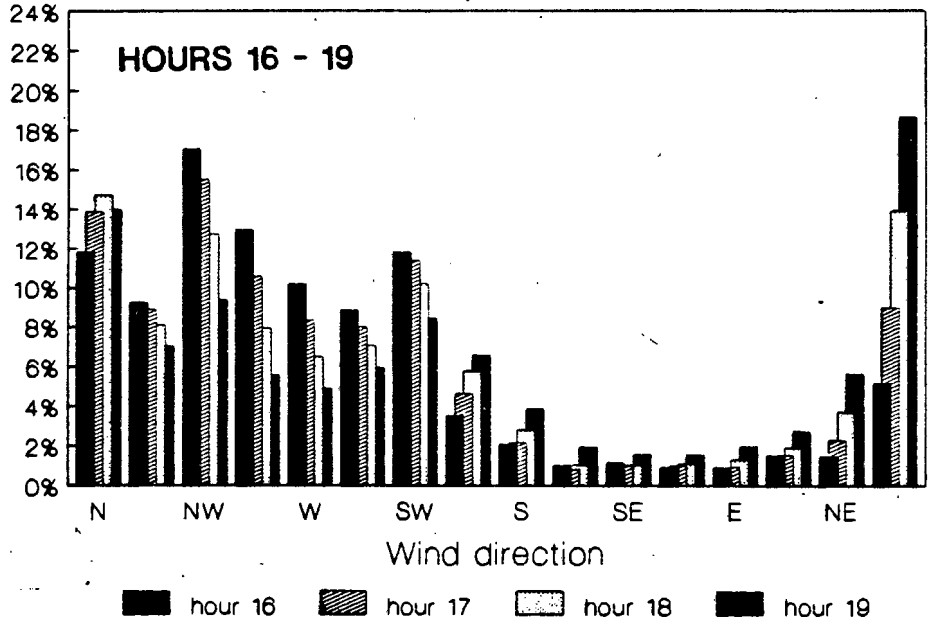

HOURS 1 - 4
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 1
hour 2
hour 3
hour 4

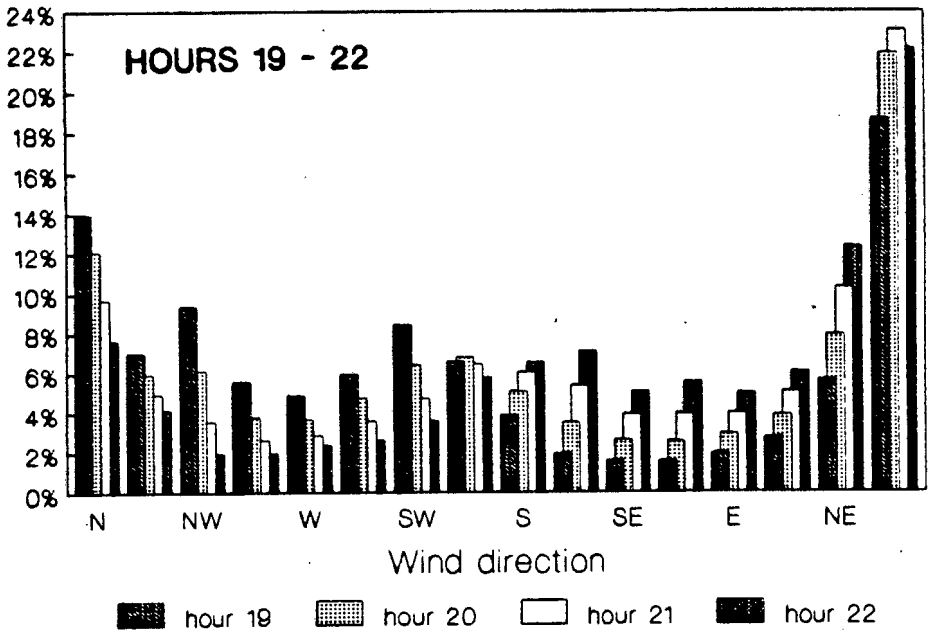

HOURS 4 - 7
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 4
hour 5
hour 6
hour 7

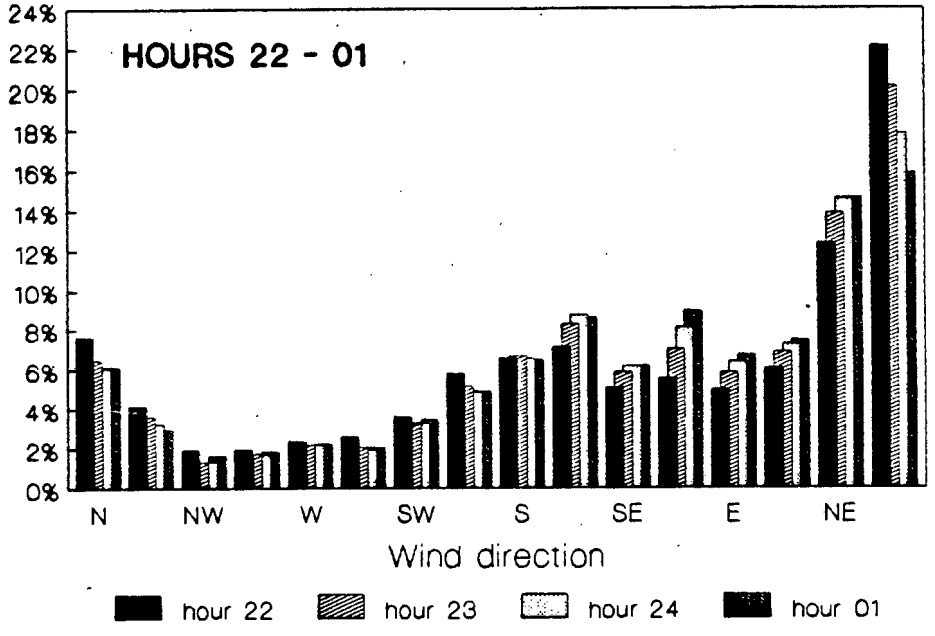

HOURS 7 - 10
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 7
hour 8
hour 9
hour 10


HOURS 10 - 13
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 10
hour 11
hour 12
hour 13


HOURS 13 - 16
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 13
hour 14
hour 15
hour 16


HOURS 16 - 19
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 16
hour 17
hour 18
hour 19


HOURS 19 - 22
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 19
hour 20
hour 21
hour 22


HOURS 22 - 01
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
N
NW
W
SW
S
SE
E
NE
Wind direction
hour 22
hour 23
hour 24
hour 01

APPENDIX A

THE NEWTON-RAPHSON METHOD

(Reference Ortega and Rheinholt (1970))

A system of n equations in n unknowns $\underline{x}$ can be represented by an n-dimensional mapping

$$F : \Re^n \rightarrow \Re^n$$

with

$$F(\underline{x}) = \underline{0}, \tag{1}$$

the rth equation of which is written

$$F_r(\underline{x}) = \underline{0}.$$

The Jacobian matrix is defined by

$$F'(\underline{x}) = \left[\frac{\partial F_r(\underline{x})}{\partial x_s}\right]_{r=1,\ldots,n;\ s=1,\ldots,n}.$$

Newton's method gives the kth iteration in the solution of (1) to be

$$\underline{x}^{(k+1)} = \underline{x}^{(k)} - [F'(\underline{x}^{(k)})]^{-1} F(\underline{x}^{(k)})$$

In practice the inverse of $F'(\underline{x}^{(k)})$ is rarely computed explicitly. Instead the system

$$F'(\underline{x}^{(k)})[\underline{x}^{(k+1)} - \underline{x}^{(k)}] = F(\underline{x}^{(k)})$$

is solved numerically for $[\underline{x}^{(k+1)} - \underline{x}^{(k)}]$. The next iteration values $\underline{x}^{(k)}$ are found by subtracting the current iteration values $\underline{x}^{(k)}$ from the vector $[\underline{x}^{(k+1)} - \underline{x}^{(k)}]$.

In our particular case we have

$$F_q(\hat{\theta}) = \sum_{i=1}^{L} \phi_{iq} \left[\exp\left(\sum_{s \in Q} \phi_{is}\hat{\theta}_s\right) - P_i\right] \quad \text{for } q \in Q$$

so that

$$\frac{\partial F_q(\hat{\theta})}{\partial \hat{\theta}_r} = \sum_{i=1}^{L} \phi_{iq}\phi_{ir} \exp\left(\sum_{s \in Q} \phi_{is}\hat{\theta}_s\right) \quad \text{for } q, r \in Q.$$

The kth iteration values are found by first solving the system

$$F'(\theta^{(k)})[\theta^{(k+1)} - \theta^{(k)}] = F(\theta^{(k)})$$

using a numerical algorithm.

Starting values. In order that the iterative process can begin it is necessary to input some starting values for the parameters. For our problem the obvious choice is to use the values of the estimated parameters in the saturated model, namely,

$$\hat{\theta}_q(\{1,\ldots,L\}) = \Sigma_i \phi_{iq} \log(P_i) \quad \text{for } q \in Q.$$

This is acceptable provided $P_i \neq 0$, in which case $\log(P_i)$ is not defined.

The problem of zero cell counts in loglinear models is one which has received attention in the literature. (See, for example, Goodman (1972), Plackett (1974) and Bishop *et al* (1975).) Many of the suggested solutions involve replacing all $\log(n_i/n_+)$ by $\log(\frac{n_i+a}{n_++b})$ with various values of a and b. This means that one *always* adjusts the observed cell counts in *all* cells, even when there are no zero cell counts. It was thought preferable only to adjust the one observed count if the count was zero. What is done is to replace $\log(n_i/n_+)$ by $\log(\epsilon/n_+)$ if (and only if) $n_i = 0$; where ϵ is some small positive number. The choice $\epsilon = 1/n$ was found to yield satisfactory starting values. •

APPENDIX B

RESULTS IN SECTION 6.1

Theorem 6.1.1. The modelling procedure for linear basis models is rotation invariant iff the basis Φ is such that,

(A) for all n $(0 \leq n \leq L-1)$ and for all $\underline{\phi}_q \in \Phi$, there exists a $\underline{\phi}_r \in \Phi$ such that

$$R^n \underline{\phi}_q = \underline{\phi}_r \qquad \text{or} \qquad R^n \underline{\phi}_q = -\underline{\phi}_r .$$

Proof. Let $\underline{\pi}, \underline{P}$ and $\Phi = \{\underline{\phi}_1, \ldots, \underline{\phi}_L\}$ be given.

Suppose that (A) holds. One must then show that

$$\underline{M}_{R^n} = R^n(\underline{M}_{R^0}) \quad \text{for} \quad n = 0, \ldots, L-1 \tag{1}$$

where $\underline{M}_{R^n}$ and $R^n(\underline{M}_{R^0})$ are defined in Section 6.1 preceeding the statement of the theorem. If (A) holds it follows that there will exist for all n and $\underline{\phi}_q \in \Phi$ a $\underline{\phi}_s \in \Phi$ such that

$$\underline{\phi}_s \sim (R^n)' \underline{\phi}_q = R^{L-n} \underline{\phi}_q . \tag{2}$$

Furthermore $\underline{\phi}_s$ will be unique (since the vectors in a basis are linearly independent). Now if (2) then

$$\underline{\phi}_s' \sim \underline{\phi}_q' R^n$$

and

$$(\underline{\phi}_q' R^n \underline{P})^2 / \hat{\text{var}}(\underline{\phi}_q' R^n \underline{P}) = (\underline{\phi}_s' \underline{P})^2 / \hat{\text{var}}(\underline{\phi}_s' \underline{P})$$

so that the estimated parameter $(\underline{\phi}_q' R^n \underline{P})$ appears in $\underline{M}_{R^n}$ iff the estimated parameter $(\underline{\phi}_s' \underline{P})$ appears in $\underline{M}_{R^0}$, i.e. $q \in Q_{R^n}$ iff $s \in Q_{R^0}$. Thus

$$\sum_{q \in Q_{R^n}} (\underline{\phi}_q' R^n \underline{P}) \underline{\phi}_q = \sum_{s \in Q_{R^0}} (\underline{\phi}_s' \underline{P}) R^n \underline{\phi}_s ,$$

i.e.

$$\underline{M}_{R^n} = R^n(\underline{M}_{R^0})$$

as was to be shown.

Having shown that (A) is sufficient for the linear modelling procedure to be rotation invariant, we now consider necessity. To achieve rotation invariance one needs to be able to write any model of the form $\underline{M}_{R^n}$ in terms of a model of the form $R^n(\underline{M}_{R^0})$ and vice-versa, i.e. one must be able to rewrite either of the two expressions shown below in the form of the other

(i)

$$\sum_q (\underline{\phi}_q' \underline{P}) R^n \underline{\phi}_q$$

(ii)

$$\sum_q (\underline{\phi}_q' R^n \underline{P}) \underline{\phi}_q .$$

Bearing in mind the possibility of one parameter models it can be seen that the following two conditions are necessary.

(B) For all $\underline{\phi}_q \in \Phi$ there exists a subset of $\{1, \ldots, L\}$, say M_q, such that

$$(\underline{\phi}_q' R^n \underline{P}) \underline{\phi}_q = \sum_{r \in M_q} (\underline{\phi}_r' \underline{P}) R^n \underline{\phi}_r .$$

(C) For all $\underline{\phi}_r \in \Phi$ there exists a subset of $\{1, \ldots, L\}$, say N_r, such that

$$(\underline{\phi}_r' \underline{P}) R^n \underline{\phi}_r = \sum_{q \in N_r} (\underline{\phi}_q' R^n \underline{P}) \underline{\phi}_q .$$

These two conditions jointly imply that for all q there exists $M_q \subseteq \{1, \ldots, L\}$ such that for all $r \in M_q$ there exists $N_r \subseteq \{1, \ldots, L\}$ such that

$$(\underline{\phi}_q' R^n \underline{P}) \underline{\phi}_q = \sum_{r \in M_q} \sum_{s \in N_r} (\underline{\phi}_s' R^n \underline{P}) \underline{\phi}_s$$

which contradicts the linear independence of the set $\{\underline{\phi}_1, \ldots, \underline{\phi}_L\}$ *unless* M_q and N_r both contain a single element; in which case (B) and (C) are equivalent to the condition:

(D) for all n and for all $\underline{\phi}_q \in \Phi$ there exists a $\underline{\phi}_r \in \Phi$ such that

$$(\underline{\phi}_q' R^n \underline{P})\underline{\phi}_q = (\underline{\phi}_r' \underline{P}) R^n \underline{\phi}_r \quad \text{for all} \quad \underline{P}.$$

It follows that (D) is a necessary condition for the procedure to be rotation invariant.

We now show that (D) is equivalent to the condition (A). Clearly (A) implies (D). To obtain the reverse implication pre-multiply both sides of the equality in (D) by $\underline{P}'(R^n)'$ to get

$$(\underline{\phi}_q' R^n \underline{P})^2 = (\underline{\phi}_r' \underline{P})^2.$$

Taking the square root on both sides gives

$$\underline{\phi}_q' R^n \underline{P} = \pm \underline{\phi}_r' \underline{P}.$$

That this holds for all $\underline{P}$ implies that

$$(R^n)' \underline{\phi}_q \sim \underline{\phi}_r,$$

and the required implication follows. •

Theorem 6.1.3. All non-zero vectors satisfying $R^n \underline{\phi} \sim \underline{\phi}$ for some n $(1 \leq n \leq L-1)$ are of one of the two forms

$$\underline{1}_d \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix} \tag{3}$$

or

$$\underline{\Lambda}_d \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix} \tag{4}$$

where

(i) in (3) d is any divisor of L, while in (4) d must be an *even* divisor of L,

(ii)

$$\underline{\Lambda}_d = \underline{1}_{d/2} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

(iii) $a_i \in \Re$ for $i = 1, \ldots, L/d$.

In order to prove the theorem the following lemma is needed.

Lemma. If $n \in \{1, \ldots, L-1\}$ then there exists a unique divisor of L , say d , with

$$n = \frac{k}{d} L \quad \text{for some integer } k, \quad 1 \leq k \leq d-1, \quad gcd(k,d) = 1$$

where $gcd(k,d)$ is the greatest common divisor (highest common factor) of k and d .

Proof. Define the set of divisors of L by

$$\mathbf{D} = \left\{ d : 2 \leq d \leq L, \ \frac{L}{d} = m \quad \text{for some natural number } m \right\}$$

and for each divisor d define

$$P_L(d) = \left\{ k\frac{L}{d} : 1 \leq k \leq d-1, \quad gcd(k,d) = 1 \right\}.$$

To prove the lemma it is sufficient to show that the $P_L(d)$ form a partition of $\{1, \ldots, L-1\}$, i.e.

(a) $\bigcup_{d \in \mathbf{D}} P_L(d) = \{1, \ldots, L-1\}$, and

(b) the $P_L(d)$ are disjoint.

To establish (a) it is sufficient to show that each of the two sets is contained in the other. Firstly then suppose that $i \in \bigcup_{d \in \mathbf{D}} P_L(d)$. From the definition of $P_L(d)$ it follows that there exists a divisor $d \in \mathbf{D}$ such that

$$i = k\frac{L}{d} \quad \text{for some integer } k, \quad 1 \leq k \leq d-1.$$

From this we reason:

(i) since d is a divisor of L, i must be an integer, and

(ii) since $1 \leq k \leq d-1$, i must satisfy $1 \leq i \leq L-1$

so that $i \in \{1, \ldots, L-1\}$.

To show the reverse inclusion suppose that $i \in \{1,\ldots,L-1\}$. Let $c = gcd(i,L)$. It then follows that there will exist two integers, say α and β which satisfy

$$i = c\alpha, \quad L = c\beta \quad \text{with} \quad 1 \leq \alpha \leq \beta - 1, \quad gcd(\alpha,\beta) = 1.$$

Hence

$$i = \left(\frac{L}{\beta}\right)\alpha \quad \text{with} \quad 1 \leq \alpha \leq \beta - 1, \quad gcd(\alpha,\beta) = 1$$

and $i \in P_L(\beta)$ with $\beta \in \mathbf{D}$, i.e.

$$i \in \bigcup_{d \in \mathbf{D}} P_L(d).$$

In order to prove (b) we begin by supposing that the $P_L(d)$ are not all disjoint. Then there will exist two divisors $d_1, d_2 \in \mathbf{D}$, $d_1 \neq d_2$ such that

$$P_L(d_1) \cap P_L(d_2) \neq \phi.$$

Take $i \in P_L(d_1) \cap P_L(d_2)$. Then

$$i = \frac{k_j}{d_j}L \quad \text{for some } k_j, \quad 1 \leq k_j \leq d_j - 1, \quad gcd(k_j,d_j) = 1 \quad \text{for } j = 1,2.$$

Hence

$$\frac{k_1}{d_1} = \frac{k_2}{d_2} \quad \text{with} \quad d_1 \neq d_2$$

which contradicts k_j and d_j being relatively prime for $j = 1$ *or* 2. •

Proof of Theorem 6.1.3. Fix n, $1 \leq n \leq L-1$. The lemma guarantees the existence of a unique divisor of L, say d, for which

$$n = \frac{k}{d}L \quad \text{for some integer } k, \quad 1 \leq k \leq d-1, \quad gcd(k,d) = 1.$$

In looking at the restriction $R^n\underline{\phi} \sim \underline{\phi}$ we need to explicitly determine the ith element of $R^n\underline{\phi}$. In fact

$$(R^n\underline{\phi})_i = \phi_{(L-n+i)\text{modulo } L}.$$

We will write $\{m\}$ as an abbreviation for m modulo L.

The two cases implicit in $R^n\underline{\phi} \sim \underline{\phi}$ are looked at separately.

(A) Suppose that $\underline{\phi} = R^n\underline{\phi}$. Equating the ith elements of the two vectors gives

$$\phi_i = \phi_{\{L-n+i\}}.$$

Now equating the $\{L-n+i\}^{\text{th}}$ elements gives

$$\begin{aligned}\phi_{\{L-n+i\}} &= \phi_{\{L-n+\{L-n+i\}\}} \\ &= \phi_{\{L-2n+i\}},\end{aligned}$$

so that one has

$$\phi_i = \phi_{\{L-n+i\}} = \phi_{\{L-2n+i\}}.$$

This process can be repeated until one "returns to" ϕ_i. This will occur for the smallest m which satisfies

$$\{L-mn\} = 0. \tag{5}$$

Now $m = d$ is a solution to (5); it is also the smallest such by its construction. Hence the restriction $\underline{\phi} = R^n\underline{\phi}$ may be written as

$$\phi_i = \phi_{\{L-n+i\}} = \cdots = \phi_{\{L-(d-1)n+i\}} \quad \text{for } i = 1,\ldots,L/d \tag{6}$$

where all the indices are distinct.

Now show that (6) is equivalent to

$$\phi_i = \phi_{i+r(L/d)} \quad \text{for } r = 1,\ldots,d-1 \tag{7}$$

which means that $\underline{\phi}$ is of the form

$$\underline{1}_d \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix}.$$

Consider two arbitrary indices in (6) which without loss may be taken to be $\{L-m_1n+i\}$ and $\{L-m_2n+i\}$ with $m_1 \neq m_2$, $0 \leq m_1, m_2 \leq d-1$. Then

$$\begin{aligned}|\{L-m_1n+i\} - \{L-m_2n+i\}| &= \{(m_2-m_1)\frac{k}{d}L\} \\ &= r\frac{L}{d}\end{aligned}$$

for some integer $r \neq 0$, where by the definition of the modulo L function

$$0 \leq r\frac{L}{d} \leq L-1.$$

Since $r \neq 0$ it follows that

$$1 \leq r \leq d-1,$$

i.e. the indices in (6) differ by $r\frac{L}{d}$ with $1 \leq r \leq d-1$. Since for each i, the $(d-1)$ indices given in (6) are distinct, it follows that (6) and (7) are equivalent.

(B) Suppose now that $\underline{\phi} = -R^n\underline{\phi}$. Going through the same process as before one obtains

$$\left.\begin{aligned}\phi_i &= (-1)\phi_{\{L-n+1\}} = (-1)^2\phi_{\{L-2n+i\}} = \cdots \\ &= (-1)^{d-1}\phi_{\{L-(d-1)n+i\}} = (-1)^d\phi_{\{L-dn+i\}} = \cdots \quad \text{for } i = 1,\ldots,L\end{aligned}\right\} \tag{8}$$

In particular

$$\phi_i = (-1)^d\phi_{\{L-dn+i\}} \quad \text{for } i = 1,\ldots,L.$$

But $\{L-dn+i\} = i$, so that if d is odd, this reads

$$\phi_i = -\phi_i \quad \text{for } i = 1,\ldots,L$$

from which it follows that $\underline{\phi} = \underline{0}$, and therefore $\underline{\phi}$ cannot appear in a basis.

Thus we need only consider even divisors d. We show that for such d

$$\phi_{(r\frac{L}{d}+i)} = (-1)^r\phi_i \quad \text{for } i = 1,\ldots,L/d. \tag{9}$$

Note firstly that

$$r\frac{L}{d} + i = L - \left(\frac{d-r}{k}\right)n + i.$$

Using (8) we can then write

$$\phi_{(r\frac{L}{d}+i)} = (-1)^{(d-r)/k}\phi_i.$$

Now d is even and k is odd (being relatively prime to the even number d) so that

$$(-1)^{(d-r)/k} = (-1)^{-r}$$

and (9) follows. From (9) (and the fact that d is even) it follows that $\underline{\phi}$ must be of the form

$$\underline{1}_{d/2} \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{L/d} \end{pmatrix}.$$

Lemma 6.1.4. (a) Let $\underline{\phi}$ have the form (3) with $L/d \geq 2$. Then $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} a_{L/d} & a_1 & a_2 & \cdots & a_{L/d-1} \\ a_{L/d-1} & a_{L/d} & a_1 & \cdots & a_{L/d-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_2 & a_3 & a_4 & \cdots & a_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L/d} \end{pmatrix} = \underline{0}. \tag{10}$$

(b) Let $\underline{\phi}$ have the form (4) with $L/d \geq 2$. Then $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} -a_{L/d} & a_1 & a_2 & \cdots & a_{L/d-1} \\ -a_{L/d-1} & -a_{L/d} & a_1 & \cdots & a_{L/d-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_2 & -a_3 & -a_4 & \cdots & a_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L/d} \end{pmatrix} = \underline{0}. \tag{11}$$

(c) Both (10) and (11) are consistent systems.

Proof. (a) Let $\underline{\phi}$ have the form (3) with $L/d \geq 2$. By the definition of $R(\underline{\phi})$, $R(\underline{\phi})$ is orthogonal iff

$$\underline{\phi}' R^n \underline{\phi} = 0 \quad \text{for } 1, 2, \ldots, L/d - 1$$

where

$$\underline{\phi}' R^n \underline{\phi} = \sum_{i=1}^{L} \phi_i \phi_{\{L-n+i\}}.$$

From the fact that for any vector in $R(\underline{\phi})$ all elements in the vector whose position differs by a multiple of L/d are equal, it follows that

$$\sum_{i=1}^{L} \phi_i \, \phi_{\{L-n+i\}} = d \sum_{i=1}^{L/d} \phi_i \, \phi_{\{L-n+i\}}.$$

Now, for $i \leq n$

$$\{L-n+i\} = L-n+i$$
$$= L/d-n+i+(d-1)L/d$$

while for $i > n$

$$\{L-n+i\} = -n+i;$$

from which it follows that for $n = 1,\ldots,L/d-1$

$$\phi_{\{L-n+i\}} = \begin{cases} \phi_{(L/d-n+i)} & \text{for } i \leq n \\ \phi_{(-n+i)} & \text{for } i > n. \end{cases} \tag{12}$$

Hence, $R(\underline{\phi})$ is orthogonal, iff

$$\sum_{i=1}^{L/d} \phi_{\{L-n+i\}}\phi_i = 0 \quad \text{for } n = 1,\ldots,L/d-1$$

where $\phi_{\{L-n+i\}}$ is given by (12). Rewriting these $(L/d-1)$ equations in matrix form gives the required result.

(b) Let $\underline{\phi}$ have the form (4) with $L/d \geq 2$. The proof of this case is similar to the above proof, the only difference being the added complication of the alternating signs.

As in the proof of (a), $R(\underline{\phi})$ is orthogonal iff

$$\sum_{i=1}^{L/d} \phi_i\,\phi_{\{L-n+1\}} = 0 \quad \text{for } n = 1,\ldots,L/d-1. \tag{13}$$

In this case however

$$\phi_{\{L-n+i\}} = \begin{cases} (-1)^{d-1}\phi(L/d-n+i) = -\phi_{(L/d-n+i)} & \text{for } i \leq n \\ \phi_{(-n+i)} & \text{for } i > n. \end{cases} \tag{14}$$

Rewriting the equations in (13) in matrix form using (14) yields the required result.

(c) A non-zero solution to both (10) and (11) is obtained by putting $a_1 = 1$ and $a_2 = \cdots = a_{L/D} = 0$. •

Proposition 6.1.5. Let $\underline{\phi}$ be a non-zero vector with the form (3). $R(\underline{\phi})$ cannot both be orthogonal to $\{\frac{1}{\sqrt{L}}\underline{1}\}$ and have its vectors orthogonal to one another.

Proof. Let $\underline{\phi}$ be a non-zero vector with the form (3). It may be assumed that $\frac{L}{d} \geq 2$.

Suppose that $R(\underline{\phi})$ is orthogonal. Then $\underline{\phi}$ satisfies (10). Pre-multiplying both sides of this equation with $\underline{1}'_L$ yields

$$\left(\sum_{\substack{j=1\\ j\neq 1}}^{L/d}\phi_j, \sum_{\substack{j=1\\ j\neq 2}}^{L/d}\phi_j, \ldots, \sum_{\substack{j=1\\ j\neq L/d}}^{L/d}\phi_j\right)\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_{L/d}\end{pmatrix} = 0$$

which can be written as

$$\sum_{i=1}^{L/d}\sum_{\substack{j=1\\ j\neq i}}^{L/d}\phi_i\phi_j = 0. \tag{15}$$

Further suppose that $R(\underline{\phi})$ is orthogonal to $\{\frac{1}{\sqrt{L}}\underline{1}_L\}$. Then

$$\sum_{i=1}^{L/d}\phi_i = 0.$$

Squaring both sides of this equation yields

$$\sum_{i=1}^{L/d}\phi_i^2 + \sum_{i=1}^{L/d}\sum_{\substack{j=1\\ j\neq i}}^{L/d}\phi_i\phi_j = 0. \tag{16}$$

Considered jointly, equations (15) and (16) imply that

$$\sum_{i=1}^{L/d}\phi_i^2 = 0$$

which implies, in turn, that $\underline{\phi} = \underline{0}$; which contradicts $\underline{\phi}$ being a non-zero vector.

•

Proposition 6.1.6. Two rotation groups of cardinalities $\frac{L}{d_1}$ and $\frac{L}{d_2}$ generated by vectors of the form (4) and satisfying the orthogonality restrictions given by (11), will be orthogonal to one another only if

$$\frac{d_1 + d_2}{gcd(d_1, d_2)} \quad \text{is odd,}$$

with no further restrictions on the elements of the generators.

Proof. Let $\underline{\phi}$ and $\underline{\psi}$ be two vectors of the form (4) whose cardinalities are $\frac{L}{d_1}$ and $\frac{L}{d_2}$ respectively. Note that the elements in $\underline{\phi}$ and $\underline{\psi}$ are repeated after every $2\frac{L}{d_1}$ and $2\frac{L}{d_2}$ elements respectively. Let

$$m = \ell cm(\frac{L}{d_1}, \frac{L}{d_2})$$

where $\ell cm(\ ,\ )$ denotes the lowest common multiple of the two arguments. Then the elements in both $\underline{\phi}$ and $\underline{\psi}$ are repeated after every $2m$ elements and L must be a multiple of $2m$.

The proof is divided into two parts:

(A) $R(\underline{\phi})$ and $R(\underline{\psi})$ are orthogonal to each other iff

(a) $\sum_{i=1}^{m} \alpha_i \beta_i = 0$ for all $\underline{\alpha} \in R(\underline{\phi}), \underline{\beta} \in R(\underline{\psi})$

or

(b) $\frac{d_1+d_2}{gcd(d_1,d_2)}$ is odd.

(B) Condition (a) cannot be satisfied by rotation groups which are themselves orthogonal.

Proof of (A). Let $\underline{\alpha}$ and $\underline{\beta}$ be elements from $R(\underline{\phi})$ and $R(\underline{\psi})$ respectively. Since the elements of both $\underline{\alpha}$ and $\underline{\beta}$ are repeated after every $2m$ elements

$$\underline{\alpha}'\underline{\beta} = \frac{L}{2m} \sum_{i=1}^{2m} \alpha_i \beta_i.$$

Using the fact that the elements within α and β are repeated with alternate signs after every $\frac{L}{d_1}$ and $\frac{L}{d_2}$ elements respectively, we may write

$$\underline{\alpha}'\underline{\beta} = \frac{L}{2m}\left[\sum_{i=1}^{m}\alpha_i\beta_i + \sum_{i=1}^{m}(-1)^{m(d_1/L)}\alpha_i\,(-1)^{m(d_2/L)}\beta_i\right]$$

$$= \frac{L}{2m}\left[1+(-1)^{(m/L)(d_1+d_2)}\right]\cdot\sum_{i=1}^{m}\alpha_i\beta_i$$

Using the result that

$$\ell cm(\frac{L}{d_1},\frac{L}{d_2}) = \frac{L}{gcd(d_1,d_2)}$$

it follows that

$$\frac{m}{L}(d_1+d_2) = \frac{d_1+d_2}{gcd(d_1,d_2)}$$

and (A) follows directly.

Proof of (B). Suppose that condition (9) holds and that $R(\underline{\phi})$ and $R(\underline{\psi})$ are orthogonal to one another. Let $\underline{\alpha}$ and $\underline{\beta}$ be arbitrary vectors in $R(\underline{\phi})$ and $R(\underline{\psi})$ respectively. In order to work with (a) we need to determine the first m elements of $\underline{\alpha}$ and $\underline{\beta}$. For $n = 0,1,\ldots,(\frac{L}{d_1}-1)$,

$$R^n\underline{\phi} = \underline{\Lambda}_{d_1}\otimes\begin{pmatrix}0 & -I_n\\ I_{(L/d_1)-n} & 0\end{pmatrix}\begin{pmatrix}\phi_1\\ \vdots\\ \phi_{L/d_1}\end{pmatrix}$$

so that the first m elements of $R^n\underline{\phi}$ are given by

$$\underline{\Lambda}_{(m(d_1/L))}\otimes\begin{pmatrix}0 & -I_n\\ I_{(L/d_1)-n} & 0\end{pmatrix}\begin{pmatrix}\phi_1\\ \vdots\\ \phi_{L/d_1}\end{pmatrix}. \tag{17}$$

Similarly the first m elements of $R^n\underline{\psi}$ are given by

$$\underline{\Lambda}_{(m(d_2/L))}\otimes\begin{pmatrix}0 & -I_n\\ I_{(L/d_2)-n} & 0\end{pmatrix}\begin{pmatrix}\psi_1\\ \vdots\\ \psi_{L/d_2}\end{pmatrix}. \tag{18}$$

Using (17) and (18) the condition (a) can be written as

$$(\underline{\Lambda}_{(m(d_1/L))}\otimes A)'(\underline{\Lambda}_{(m(d_2/L))}\otimes B) = 0 \tag{19}$$

where

$$A = \begin{pmatrix} \phi_1 & -\phi_{L/d_1} & \cdots & -\phi_2 \\ \phi_2 & \phi_1 & \cdots & -\phi_3 \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{L/d_1} & \phi_{(L/d_1)-1} & \cdots & \phi_1 \end{pmatrix}$$

$$B = \begin{pmatrix} \psi_1 & -\psi_{L/d_2} & \cdots & -\psi_2 \\ \psi_2 & \psi_1 & \cdots & -\psi_3 \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{L/d_2} & \psi_{(L/d_2)-1} & \cdots & \psi_1 \end{pmatrix}.$$

Pre-multiplying both sides of (19) with

$$(\phi_{L/d_1},\ \phi_{(L/d_1)-1}, \ldots, \phi_1)$$

and post-multiplying both sides with

$$(\psi_{L/d_2},\ \psi_{(L/d_2)-1}, \ldots, \psi_1)'$$

yields, using Lemma 6.1.4(b),

$$\left(\underline{\Lambda}_{(m(d_1/L))} \otimes \begin{pmatrix} 0 \\ \vdots \\ 0 \\ c_1 \end{pmatrix}\right)' \left(\underline{\Lambda}_{(m(d_2/L))} \otimes \begin{pmatrix} 0 \\ \vdots \\ 0 \\ c_2 \end{pmatrix}\right) = 0 \quad \text{with } c_1, c_2 > 0. \qquad (20)$$

The ith vector $(i = 1, 2)$ on the left side of (20) has non-zero elements only in every position which is a multiple of L/d_i . The first position at which *both* have non-zero elements is thus at the mth position. Thus, since the last element in $\Lambda_{(m(d_i/L))}$ $(i = 1, 2)$ is 1 and not -1, (20) reads as

$$c_1 c_2 = 0 \quad \text{with } c_1, c_2 > 0$$

which demonstrates the required contradition. •

Proposition 6.1.7. The maximum number of vectors available for simultaneous inclusion in a rotation invariant model basis is less than or equal L with equality holding only if $L = 2^m$ for some natural number m.

Proof. L is taken to be even. Let the prime decomposition of L be

$$2^m P_1^{m_1} \cdots P_k^{m_k}$$

where the P_i are distinct *odd* primes and m and the m_i are positive integers. Any even divisor of L is then of the form

$$2^n P_1^{n_1} \cdots P_k^{n_k}$$

where $1 \leq n \leq m$ and $0 \leq n_i \leq m_i$ for $i = 1, \ldots, k$, and gives rise to a rotation group of cardinality

$$\frac{L}{2^n P_1^{n_1} \cdots P_k^{n_k}}. \tag{21}$$

From proposition 6.1.6 it is known that two rotation groups of the same cardinality cannot appear in the same rotation invariant basis, and hence we can restrict attention to rotation groups of distinct cardinalities. For each fixed n, let $C(n)$ denote the set of (distinct) cardinalities of the form (21) obtained by varying the n_i $(i = 1, \ldots, k)$. We show that

(A) for each n only one of the rotation groups with cardinalities in $C(n)$ may be included in any model basis. For each n, the natural choice for this rotation group is that with the largest cardinality, i.e. with cardinality that of the largest element in $C(n)$, namely $L/2^n$.

We then show that

(B) the set of rotation groups of cardinality $L/2^n$ for $n = 1, \ldots, m$ *can* coexist in the same basis.

This means that, including $\frac{1}{\sqrt{L}} 1_L$, the maximum number of vectors available

for simultaneous inclusion in a rotation invariant basis, is

$$1 + L\sum_{n=1}^{m} \frac{1}{2^n}$$
$$= 1 + L(1 - \frac{1}{2^m})$$
$$= L + (1 - P_1^{m_1} \cdots P_k^{m_k})$$

and the proposition follows directly.

Proof of A. Fix n. Let L/d_1 and L/d_2 be two elements of $C(n)$, so that

$$d_1 = 2^n P_1^{r_1} \cdots P_k^{r_k}$$
$$d_2 = 2^n P_1^{s_1} \cdots P_k^{s_k}$$

for some r_i, s_i ; $1 \leq r_i, s_i \leq m_i$ for $i = 1, \ldots k$. Then

$$gcd(d_1, d_2) = 2^n P_1^{t_1} \cdots P_k^{t_k} \quad \text{where } t_i = \min(r_i, s_i)$$

and

$$\frac{d_1 + d_2}{gcd(d_1, d_2)} = P_1^{r_1 - t_1} \cdots P_k^{r_k - t_k} + P_1^{s_1} \cdots P_k^{s_k - t_k} \tag{22}$$

In (22) each of the P_i is odd. Now any positive power of an odd number is odd and the product of two odd numbers is odd. Thus the two terms on the left side of (22) are both odd; and their sum is even. Thus by proposition 6.1.6 the corresponding rotation groups cannot appear simultaneously in any model basis.

Proof of B. Consider two rotation groups of cardinality $L/2^{n_1}$ and $L/2^{n_2}$. Suppose that $n_1 > n_2$. Then

$$\frac{2^{n_1} + 2^{n_2}}{gcd(2^{n_1}, 2^{n_2})} = 2^{n_1 - n_2} + 1$$

which is odd, so that by proposition 6.1.6 the associated rotation groups will be orthogonal to each other and thus can be placed in the same model basis. •

Proposition 6.1.8. Let $L = 2^m$ and let

$$\underline{\phi} = \underline{\Lambda}_{2^{m-s}} \otimes \begin{pmatrix} a_1 \\ \vdots \\ a_{2^s} \end{pmatrix} \quad \text{for some } s, \quad 0 \leq s \leq m-1 .$$

(a) For $s = 0$ and 1; $R(\underline{\phi})$ is orthogonal.

(b) For $s = 2, \ldots, m-1$; $R(\underline{\phi})$ is orthogonal iff

$$\begin{pmatrix} -a_{2^s} & a_1 & a_2 & \cdots & a_{2^s-1} \\ -a_{2^s-1} & -a_{2^s} & a_1 & \cdots & a_{2^s-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{\frac{1}{2}2^s+2} & -a_{\frac{1}{2}2^s+3} & -a_{\frac{1}{2}2^s+4} & \cdots & -a_{\frac{1}{2}2^s+1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{2^s} \end{pmatrix} = \underline{0} \qquad (23)$$

which is a consistent system.

Proof. (a) For $s = 0$, $\underline{\phi} = \underline{\Lambda}_{2^m}$ which has cardinality 1, so that $R(\underline{\phi})$ is (trivially) orthogonal. For $s = 1$

$$\underline{\phi} = \underline{\Lambda}_{2^{m-1}} \otimes \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \text{for some } a_1, a_2 \in \Re$$

which has cardinality 2, and

$$\begin{aligned} \underline{\phi}'(R^1\underline{\phi}) &= (\underline{\Lambda}'_{2^{m-1}} \otimes (a_1, a_2))\left(\underline{\Lambda}_{2^{m-1}} \otimes \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right) \\ &= 2^{m-1}(-a_1a_2 + a_1a_2) \\ &= 0 \end{aligned}$$

so that $R(\underline{\phi})$ is orthogonal.

(b) Let $2 \leq s \leq m-1$. From the proof of lemma 6.1.5, $R(\underline{\phi})$ is orthogonal iff

$$\underline{\phi}' R^n \underline{\phi} = 0 \quad \text{for } n = 1, \ldots, 2^s - 1.$$

We begin by showing that the first $\frac{1}{2}2^s$ of these constraints are sufficient – in the sense that the remaining constraints are just repetitions of these.

Let n satisfy

$$\frac{1}{2}2^s \leq n \leq 2^s - 1.$$

Then n can be written as

$$2^s - m \quad \text{where} \quad 1 \leq m \leq \frac{1}{2}2^s$$

and

$$\underline{\phi}' R^n \underline{\phi} = \underline{\phi}' R^{2^s - m} \underline{\phi} = \underline{\phi}' (R^m)' (R^{2^s} \underline{\phi}).$$

By the construction of $\underline{\phi}$, $R^{2^s}\underline{\phi} = -\underline{\phi}$. Thus

$$\underline{\phi}' R^n \underline{\phi} = -\underline{\phi}' (R^m)' \underline{\phi}$$

the left side of which, by a well known property of quadratic forms can be written as $-\underline{\phi}' R^m \underline{\phi}$. Thus for $n = \frac{1}{2}2^s, \ldots, 2^s - 1$

$$\underline{\phi}' R^n \underline{\phi} = -\underline{\phi}' R^m \underline{\phi} \quad \text{for some } m, \quad 1 \leq m \leq \frac{1}{2}2^s,$$

and $R(\underline{\phi})$ is orthogonal iff

$$\underline{\phi}' R^n \underline{\phi} = 0 \quad \text{for } n = 1, \ldots, \frac{1}{2}2^s.$$

Now show that the last of these equations is redundant: for $n = \frac{1}{2}2^s$

$$\begin{aligned} \underline{\phi}' R^n \underline{\phi} &= \underline{\phi}' R^{2^s - \frac{1}{2}2^s} \underline{\phi} \\ &= -\underline{\phi}' R^{\frac{1}{2}2^s} \underline{\phi} = -\underline{\phi}' R^n \underline{\phi} \end{aligned}$$

which implies that

$$\underline{\phi}' R^n \underline{\phi} = 0.$$

Thus $R(\underline{\phi})$ is orthogonal iff

$$\underline{\phi}' R^n \underline{\phi} = 0 \quad \text{for } n = 1, \ldots, \frac{1}{2}2^s - 1. \tag{24}$$

Using an argument similar to that given in the proof of Lemma 6.1.5, the system (24) can be written in matrix form as (23). •

APPENDIX C

RESULTS IN SECTION 6.3

Theorem 6.3.1. The pairwise modelling procedure is rotation invariant iff the basis Φ is such that

(A) for all n $(1 \leq n \leq L)$ and for all $q \in \{1, \ldots, \frac{1}{2}(L-1)\}$, there exists an $r \in \{1, \ldots, \frac{1}{2}(L-1)\}$ such that

$$\begin{aligned}&\phi_{i,2r}\phi_{j,2r} + \phi_{i,2r+1}\phi_{j,2r+1}\\ &= (R^n \underline{\phi}_{2q})_i (R^n \underline{\phi}_{2q})_j + (R^n \underline{\phi}_{2q+1})_i (R^n \underline{\phi}_{2q+1})_j \quad \text{for all } i, j.\end{aligned}$$

Proof. Put $\mathbf{S} = \{1, \ldots, \frac{1}{2}(L-1)\}$. By an argument similar to that used in the proof of Theorem 6.1.1, it can be shown that rotation invariance of the pairwise modelling procedure is equivalent to

(B) for all n and for all $q \in \mathbf{S}$, there exists an $r \in \mathbf{S}$, such that

$$\begin{aligned}&(\underline{\phi}'_{2r}\underline{P})R^n\underline{\phi}_{2r} + (\underline{\phi}'_{2r+1}\underline{P})R^n\underline{\phi}_{2r+1}\\ &\qquad = (\underline{\phi}'_{2q}R^n\underline{P})\underline{\phi}_{2q} + (\underline{\phi}'_{2q+1}R^n\underline{P})\underline{\phi}_{2q+1}\end{aligned}$$

for all vectors of sample proportions $\underline{P}$.

Suppose that the pairwise modelling procedure is rotation invariant so that (B) holds. Pre-multiplying the expression in (B) by $\underline{P}'R^{-n}$ throughout, gives

$$(\underline{\phi}'_{2r}\underline{P})^2 + (\underline{\phi}'_{2r+1}\underline{P})^2 = (\underline{\phi}'_{2q}R^n\underline{P})^2 + (\underline{\phi}'_{2q+1}R^n\underline{P})^2$$

Expanding each of the terms in this equation, using the fact that

$$(\underline{\phi}'_p R^n \underline{P}) = (R^{L-n}\underline{\phi}_p)'\underline{P}$$

gives

$$\begin{aligned}&\Sigma_i\Sigma_j\{\phi_{i,2r}\phi_{j,2r} + \phi_{i,2r+1}\phi_{j,2r+1}\}P_iP_j\\ &=\Sigma_i\Sigma_j\{(R^{L-n}\underline{\phi}_{2q})_i (R^{L-n}\underline{\phi}_{2q})_j + (R^{L-n}\underline{\phi}_{2q+1})_i (R^{L-n}\underline{\phi}_{2q+1})_j\}P_iP_j.\end{aligned}$$

Since this holds for all $\underline{P}$ and since $(L-n)$ transverses $\{1,\ldots,L\}$ as n runs from 1 to L, (A) follows.

For the reverse implication, suppose that (A) holds. Then

(C) for all n $(1 \le n \le L)$ and for all $q \in \mathbf{S}$, there exists an $r \in \mathbf{S}$ such that

$$\phi_{i,2r}\phi_{j,2r} + \phi_{i,2r+1}\phi_{j,2r+1}$$
$$= (R^{L-n}\underline{\phi}_{2q})_i (R^{L-n}\underline{\phi}_{2q})_j + (R^{L-n}\underline{\phi}_{2q+1})_i (R^{L-n}\underline{\phi}_{2q+1})_j \quad \text{for all } i,j.$$

In this equation multiply through by P_i and then sum over all i, to get for all j

$$(\Sigma_i \phi_{i,2r} P_i)\, \phi_{j,2r} + (\Sigma_i \phi_{i,2r+1} P_i)\phi_{j,2r+1}$$
$$= (\Sigma_i (R^{L-n}\underline{\phi}_{2q})_i P_i)\, (R^{L-n}\underline{\phi}_{2q})_j + (\Sigma_i (R^{L-n}\underline{\phi}_{2q+1})_i P_i)\, (R^{L-n}\underline{\phi}_{2q+1})_j$$

which, in turn, implies, that

$$(\underline{\phi}'_{2r}\underline{P})\underline{\phi}_{2r} + (\underline{\phi}'_{2r+1}\underline{P})\underline{\phi}_{2r+1}$$
$$= (\underline{\phi}'_{2q} R^n \underline{P}) R^{L-n}\underline{\phi}_{2q} + (\underline{\phi}'_{2q+1} R^n \underline{P}) R^{L-n}\underline{\phi}_{2q+1}.$$

Multiplying this equation through by R^n yields (B). •

Multiplication of elements in $\Re^2$

Let $\alpha = (\alpha_1;\alpha_2)$ and $\beta = (\beta_1;\beta_2)$ be two elements in $\Re^2$. Then the definition of the product of α and β is

$$\alpha\beta = (\alpha_1\beta_1 - \alpha_2\beta_2;\ \alpha_1\beta_1 + \alpha_2\beta_2).$$

One has that

$$|\alpha\beta| = |\alpha||\beta|$$

and

$$\mathrm{Arg}(\alpha\beta) = \mathrm{Arg}\,\alpha + \mathrm{Arg}\,\beta + 2k\pi \quad \text{for some integer } k.$$

(The definition of Arg is given in Section 3 just below proposition 6.3.2.)

Proposition 6.3.2. A basis Φ (whose elements have been formed into the pairs z_{iq} as in (6.3.4)) is pairwise rotation invariant iff

(D) for each $q = 1, \ldots, \frac{1}{2}(L-1)$ there exists an $\alpha \in \Re^2$, $|\alpha| = 1$ such that

$$\alpha z_{iq} = z_{\{i+1\},q} \quad \text{for } i = 1, \ldots, L.$$

Proof. Let z_{iq} be written in terms of its polar co-ordinates as in (6.3.6), i.e.

$$\begin{aligned} z_{iq} &= (\phi_{i,2q};\ \phi_{i,2q+1}) \\ &= \sqrt{\frac{2}{L}}(\cos \text{ Arg } z_{iq};\ \sin \text{ Arg } z_{iq}). \end{aligned}$$

Suppose that Φ is pairwise rotation invariant. From the definition and the proof of Theorem 6.3.1, for each q and n

$$\begin{aligned} &\phi_{i,2q}\,\phi_{j,2q} + \phi_{i,2q+1}\,\phi_{j,2q+1} \\ &= \phi_{\{i+n\},2q}\,\phi_{\{j+n\},2q} + \phi_{\{i+n\},2q+1}\,\phi_{\{j+n\},2q+1} \quad \text{for all } i, j. \end{aligned}$$

For each q and n this equation can be re-expressed in terms of the polar co-ordinates of z_{iq} and $z_{\{i+n\},q}$. Using the fact that for any two angles θ_1 and θ_2

$$\cos(\theta_1 - \theta_2) = \cos\theta_1 \cos\theta_2 + \sin\theta_1 \sin\theta_2 \tag{25}$$

the resulting equation simplifies to

$$\cos(\text{Arg } z_{iq} - \text{Arg } z_{jq}) = \cos(\text{Arg } z_{\{i+n\},q} - \text{Arg } z_{\{j+n\},q}) \quad \text{for all } i, j. \tag{26}$$

Now consider three points z_{iq}; $z_{\{i+1\},q}$ and $z_{\{i+2\},q}$ which all lie on the same circle. There will always exist $\alpha, \beta \in \Re^2$, $|\alpha| = |\beta| = 1$ such that

$$\alpha z_{iq} = z_{\{i+1\},q}$$

$$\beta z_{\{i+1\},q} = z_{\{i+2\},q}.$$

Using standard proportions of multiplication in $\Re^2$, it follows from these two equations that

$$\cos(\text{Arg } z_{iq} - \text{Arg } z_{\{i+1\},q}) = \cos(\text{Arg } \alpha)$$

and

$$\cos(\text{Arg } z_{\{i+1\},q} - \text{Arg } z_{\{i+2\},q}) = \cos(\text{Arg } \beta).$$

From (25) it then follows that $\text{Arg}\,\alpha = \text{Arg}\,\beta$ which means that $\alpha = \beta$. Hence there exists a single element of $\Re^2$ which will take you from any z_{iq} to the next one, and (D) is proved.

For the reverse implication, note that from (D) it follows for all i and j

$$|\text{Arg}\, z_{iq} - \text{Arg}\, z_{jq}| = |i-j|\text{Arg}\,\alpha + k2\pi \quad \text{for some integer } k$$

which clearly implies that (26) holds for all n, (which is equivalent to the basis being pairwise rotation invariant). •

Lemma 6.3.3. A basis Φ is pairwise rotation invariant iff for $q = 1, \ldots, \frac{1}{2}(L-1)$

(E) $\text{Arg}\, z_{iq} = \{a(i-1)+b\}\frac{2\pi}{L}$ for $i = 1, \ldots, L$ for some $a, b \in \Re$.

Proof. Suppose that the basis is pairwise rotation invariant. Then (D) of the previous proposition holds. Let a and b be chosen such that

$$\text{Arg}\,\alpha = a\frac{2\pi}{L}, \quad \text{Arg}\, z_{iq} = b\frac{2\pi}{L} \quad \text{with } 0 \le a, b < L.$$

From (D)

$$\begin{aligned} \text{Arg}\, z_{\{i+1\},q} &= \text{Arg}\, z_{iq} + \text{Arg}\,\alpha + k2\pi \quad \text{for some integer } k \\ &= \text{Arg}\, z_{iq} + a\frac{2\pi}{L} + k2\pi. \end{aligned} \tag{27}$$

This is used to prove (E) by induction. For $i = 1$, it follows from (27) that

$$\begin{aligned} \text{Arg}\, z_{2q} &= (a+b)\frac{2\pi}{L} + k2\pi \\ &= \{a+b\}\frac{2\pi}{L}. \end{aligned}$$

For $1 < i \le L$ suppose that

$$\text{Arg}\, z_{iq} = \{a(i-1)+b\}\frac{2\pi}{L} \tag{28}$$

Then from (27) and (28) it follows that

$$\begin{aligned} \text{Arg}\, z_{\{i+1\},q} &= \{a(i-1)+b\}\frac{2\pi}{L} + a\frac{2\pi}{L} + k2\pi \\ &= \{ai+b\}\frac{2\pi}{L}. \end{aligned}$$

Hence (E) follows by the induction principle.

Now consider the reverse implication. Suppose that (E) holds. We show that the basis is pairwise rotation invariant by showing that (D) holds. To prove (D) we demonstrate the existence of an α with the required properties. Put

$$\alpha = (\cos a\frac{2\pi}{L}; \sin a\frac{2\pi}{L}).$$

Then clearly $|\alpha| = 1$. Furthermore from (E) it follows that for $i = 1, \ldots, L$

$$\begin{aligned} \text{Arg}\,(\alpha z_{iq}) &= a\frac{2\pi}{L} + \{a(i-1)+b\}\frac{2\pi}{L} + k2\pi \\ &= \{ai+b\}\frac{2\pi}{L} \\ &= \text{Arg}\, z_{\{i+1\},q}. \end{aligned}$$

In addition, since all the z_{iq} $(i = 1, ..., L)$ have equal length and since $|\alpha| = 1$,

$$|\alpha z_{iq}| \quad (= |z_{iq}|) = |z_{\{i+1\},q}|.$$

That is, αz_{iq} and $z_{\{i+1\},q}$ have the same argument and the same length; hence they are equal. i.e. $|\alpha| = 1$ and $\alpha z_{iq} = z_{\{i+1\},q}$ for $i = 1, \ldots, L$. •

Lemma 6.3.4. For all $a, b \in \Re$

$$\begin{aligned} &\left[\cos((ai+b)\frac{2\pi}{L}); \quad \sin((ai+b)\frac{2\pi}{L})\right]_{i=0,\ldots,L-1} \\ &\qquad \approx \left[\cos(ai\frac{2\pi}{L}); \quad \sin(ai\frac{2\pi}{L})\right]_{i=0,\ldots,L-1}. \end{aligned}$$

Proof. Put

$$\begin{aligned} T(a,b) = &\{\cos((ai+b)\frac{2\pi}{L}) \cdot \cos((aj+b)\frac{2\pi}{L})\} \\ &+ \{\sin((ai+b)\frac{2\pi}{L}) \cdot \sin((aj+b)\frac{2\pi}{L}). \end{aligned}$$

It is sufficient to show that for all $a, b \in \Re$ that $T(a,b) = T(a,0)$. This follows immediately from (25). •

Lemma 6.3.5. Let $0 < a < L$. Then the two conditions

$$(I) \qquad \sum_{i=0}^{L-1} \cos(ai\frac{2\pi}{L}) = 0$$

$$(II) \qquad \sum_{i=0}^{L-1} \sin(ai\frac{2\pi}{L}) = 0$$

are satisfied simultaneously iff $a \in \{1, \ldots, L-1\}$.

Proof. From the list of identities given in Bloomfield (1976, p15) one can establish that for $0 \leq a < L$:

$$\sum_{i=0}^{L-1} \cos(ai\frac{2\pi}{L}) = \begin{cases} \cos(\frac{L-1}{L}a\pi)(\sin a\pi)/\sin(a\frac{\pi}{L}) & \text{provided } a \neq 0 \\ L & \text{if } a = 0 \end{cases}$$

$$\sum_{i=0}^{L-1} \sin(ai\frac{2\pi}{L}) = \begin{cases} \sin(\frac{L-1}{L}a\pi)(\sin a\pi)/\sin(a\frac{\pi}{L}) & \text{provided } a \neq 0 \\ 0 & \text{if } a = 0. \end{cases}$$

It follows that (I) holds if

$$a \in \{1, \ldots, L-1\} \quad \text{or} \quad a = \frac{L}{L-1}\frac{1}{2} \quad \text{or} \quad a = \frac{L}{L-1}\frac{3}{2}$$

while (II) holds if

$$a \in \{a, \ldots, L-1\} \quad \text{or} \quad a = \frac{L}{L-1}.$$

Thus (I) and (II) are satisfied simultaneously iff

$$a \in \{1, \ldots, L-1\}. \quad \bullet$$

REFERENCES

BATSCHELET, E. (1981). *Circular statistics in biology*, Academic Press, New York.

BISHOP, Y., FIENBERG, S. and HOLLAND, P. (1975). *Discrete multivariate analysis : theory and practice*, MIT Press, New York.

BLOOMFIELD, P. (1976). *Fourier analysis of time series : an introduction*, Wiley, New York.

FIENBERG, S. (1977). *The analysis of cross-classified categorical data*, MIT Press, New York.

GOODMAN, L. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association*, **65**, 226-256.

GOODMAN, L. (1972). Some multiplicative models for the analysis of cross-classified data. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* edited by L. Le Cam *et al.*, **1**, 649-696, Berkeley, University of California Press.

HALL, P. (1983). Orthogonal series methods for both qualitative and quantitative data. *The Annals of Statistics*, **11**, 1156-1174.

HEWLETT, D. and PLACKETT, R. (1950). Statistical aspects of the independent joint action of poisons, particularly insecticides II. Examination of data for agreement with the hypothesis. *The Annals of Applied Biology*, **37**, 527-52.

Gloss (1986) ?

KRONMAL, R. and TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association*, **71**, 391-399.

LIANG, W. and KRISNAIAH, P.R. (1985a). Nonparametric iterative estimation of multivariate binary density. *Journal of Multivariate Analysis*, **16**, 162-172.

DIGGLE, P. and HALL, P. (1986). The selection of terms in an orthogonal series density estimator. Journal of the American Statistical Association, 81, 230-233.

LIANG, W. and KRISNAIAH, P.R. (1985b). Multi-stage nonparametric estimation of density function using orthonormal systems. *Journal of Multivariate Analysis*, **17**, 228-241.

LINHART, H. and ZUCCHINI, W. (1986a). *Model selection*, Wiley, New York.

LINHART, H. and ZUCCHINI, W. (1986b). Finite sample selection criteria for multinomial models. *Statistische Hefte*, **27**, 173-178.

MULLER, T. and MAYHALL, J. (1971). Analysis of contingency table data on *torus mandibularis* using a loglinear model. *American Journal of Physical Anthropology*, **34**, 149-154.

ORTEGA, J.H. and RHEINHOLT, W.C. (1970). *Iterative solutions of nonlinear equations in several variables.* Academic Press, New York.

OTT, J. and KRONMAL, R. (1976). Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association*, **71**, 391-399.

PLACKETT, R. (1974). *The analysis of categorical data*, Griffin, London.

RAKTOE, B., HEDAYAT, A. and FEDERER, W. (1981). *Factorial designs*, Wiley, New York.

SCHOENER, T. (1968). The *anolis* lizards of Bimini : resource partitioning in complex fauna. *Ecology*, **49**, 707-726.

STOUFER, S., SUCHMAN, E., DEVINNEY, L., STAR, S. and WILLIAMS, R. (1949). *The American soldier*, **1**, Princeton University Press, Princeton, N.J.

STUART, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, **40**, 105-110.