# COHERENT PREDICTIVE PROBABILITIES

A THESIS
SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY IN STATISTICS
IN THE
UNIVERSITY OF CANTERBURY
by

Andrea Robyn Piesse

University of Canterbury
1996

# Abstract

The main aim of this thesis is to study ways of predicting the outcome of a vector of category counts from a particular group, in the presence of like data from other groups regarded exchangeably with this one and with each other. The situation is formulated using the subjectivist framework and strategies for estimating these predictive probabilities are presented and analysed with regard to their coherency. The range of estimation procedures considered covers naive, empirical Bayes and hierarchical Bayesian methods. Surprisingly, it turns out that some of these strategies must be asserted with zero probability of being used, in order for them to be coherent. A theory is developed which proves to be very useful in determining whether or not this is the case for a given collection of predictive probabilities. The conclusion is that truly Bayesian inference may lie behind all of the coherent strategies discovered, even when they are proposed under the guise of some other motivation.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Many problems in statistics, and especially in biostatistics, involve the evaluation or comparison of predicted values. Geisser [32, 33] has argued that the long neglected predictive approach to statistical inference is, therefore, often more appropriate than the classical estimative approach. This point of view is accepted and adopted in the work that follows. Hence, the main focus is on the prediction of future observables, as opposed to the more common, if less securely founded, emphasis on the estimation and testing of unknown parameters.

We begin by introducing the context in which the estimation of predictive probabilities is to be studied, and some relevant notation.

## 1.1 Formulation of the Problem

Suppose we have a sample of $r$ items from each of $N$ groups, and that each item can be classified into one of $K+1$ distinct categories. Let

$$\underline{Y}^{(i)} = \begin{pmatrix} Y_1^{(i)} \\ Y_2^{(i)} \\ \vdots \\ Y_{K+1}^{(i)} \end{pmatrix}$$

be the vector of category counts for the $i^{\text{th}}$ group, $i = 1, \ldots, N$, where $\sum_{j=1}^{K+1} Y_j^{(i)} = r$. The number of possible outcomes for $\underline{Y}^{(i)}$ is the number of ways of putting $r$ things into $K+1$ boxes, namely ${}^{r+K}C_K$.

**Definition 1.1.1** *The* ${}^{r+K}C_K$ *possible outcomes for each* $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, *will be referred to as* ***types.***

It will be convenient to assign an ordering to the types. Suppose type $s$ is

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{K+1} \end{pmatrix}$$

and type $t$ is

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{K+1} \end{pmatrix}.$$

Let $J$ be the smallest $j \in \{1, \ldots, K+1\}$ such that $a_j \neq b_j$. Type $s$ precedes type $t$ in order (where it is understood that $s < t$) if and only if $a_J > b_J$. Hence, for example, type 1 is

$$\begin{pmatrix} r \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and type ${}^{r+K}C_K$ is

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ r \end{pmatrix}.$$

Suppose these $N$ groups are regarded to be of a similar nature, *and* there is another group of interest that is regarded to be similar with them. Often in such a situation we would like to make inferences about this extra group, based on the observations from the other $N$ groups. In particular, it would be useful to estimate the predictive probabilities

$$P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right), \tag{1.1}$$

where $\underline{Y}^{(N+1)}$ is the vector of category counts for the extra group and $\underline{y}$ is a given type.

As a concrete example, imagine samples of $r$ apples from each of $N$ orchards, where every apple has been classified into one of $K+1$ categories depending on its quality. Suppose these $N$ orchards are similar with regard to their age, location, climate, soil conditions, fertilisation and pruning procedures and the apple variety grown. The owner of a new orchard that is similar in all respects to these others, except for age, may use this sample data to predict what results might be expected from her own orchard when it does reach the same stage of maturity.

The aim of this thesis is to investigate various methods and strategies for specifying the probabilities in (1.1) and to analyse their implications under a subjectivist framework.

Chapter 2 reviews subjective statistics and reformulates the problem in this context, while Chapters 3, 4 and 5 look at probability estimation from a naive, empirical Bayes and hierarchical Bayesian point of view, respectively. Chapter 6 summarises the findings of this study and suggests directions for future research. Appendix A contains listings of computer programs that were written to produce some of the results embodied in this thesis, while in Appendix B appears a data set referred to in the text.

This manuscript was typeset using the LaTeX document preparation system.

# Chapter 2

# Subjective Statistics

The word 'probability' is constantly in use in empirical science, mathematics, statistics, philosophy and everyday life. Not surprisingly, therefore, it has acquired a number of different meanings and these may be classified into essentially three classes. The frequentist theory (formulated in Venn [87]) identifies probability with the limit of a relative frequency of occurrence of an event in a large number of presume independent repetitions of a given situation. Despite arguments that such a frequency concept rarely, if ever, applies, this theory is still very widely held and dominates much of the current statistical practice and education. An alternative view is the logical formulation (made explicit in Keynes [53]) which takes probability as representing a logical relation between a proposition and a body of knowledge, *i.e.*, between one statement and another (set of) statement(s) representing evidence. On the contrary, the subjective formulation proposes probability to be the measure of a person's degree of belief in the possible occurrence of an event. Though prevalent among the early emerging ideas of probability in Europe, the operational subjective theory of statistics has developed slowly, and in the hands of a minority, over the past two centuries. Significant contributions have come from De Morgan [25], Borel [10], Ramsey [76], Savage [81] and particularly the writings of de Finetti [22, 23, 24], whose approach is outlined in this chapter and whose characterization of the concept of coherency underlies the entire thesis.

## 2.1 Subjective Statistical Methods

To introduce the language and notation of the subjectivist framework a number of definitions, taken in the main from Lad [56], are now given.

**Definition 2.1.1** *An **operationally defined measurement** is a specified procedure of action which, when followed, yields a number. The number yielded by performing an operationally defined measurement is called a **quantity**.*

**Definition 2.1.2** *The set of all numbers that can possibly result from performing the operational measurement procedure that defines a quantity, $X$, is called the **realm** of the quantity, denoted by $\mathcal{R}(X)$.*

**Definition 2.1.3** *If $E$ is a quantity for which $\mathcal{R}(E) = \{0, 1\}$, then $E$ is also called an **event**.*

**Event Notation**: The notation of parentheses around an arithmetic statement, such as $(X = x_i)$, is used to denote an *event* defined to equal 1 if the expression is true and to equal 0 if it is false.

**Definition 2.1.4** *$N$ events constitute a **partition** if at most one of them can occur and it is impossible that none of them occurs. Their sum necessarily equals 1: $\sum_{i=1}^{N} E_i = 1$. Each of the events that together constitute a partition is called a **constituent** of the partition.*

More generally, consider the following.

**Definition 2.1.5** *A **vector of quantities**, $\underline{X}_N$, is a vector each of whose components is a quantity.*

**Definition 2.1.6** *The **realm of a vector of quantities**, $\mathcal{R}(\underline{X}_N)$, is the set of all vectors that can possibly result from the performance of all their operational definitions.*

**Definition 2.1.7** *Consider a vector of quantities, $\underline{X}_N$, whose realm is denoted by $\mathcal{R}(\underline{X}_N)$. The **realm matrix** for $\underline{X}_N$, denoted by $\boldsymbol{R}(\underline{X}_N)$, is the matrix whose columns are the vector elements listed in $\mathcal{R}(\underline{X}_N)$.*

If $\underline{X}_N$ is a vector of quantities with realm matrix $\boldsymbol{R}(\underline{X}_N) = [\underline{x}_{N1}\,\underline{x}_{N2}\cdots\underline{x}_{Nz}]$, then the components of the event vector $\underline{Q}(\underline{X}_N) = [(\underline{X}_N = \underline{x}_{N1}), \ldots, (\underline{X}_N = \underline{x}_{Nz})]^T$ constitute the partition generated by $\underline{X}_N$. The equation

$$\underline{X}_N = \boldsymbol{R}(\underline{X}_N).\underline{Q}(\underline{X}_N) \tag{2.1}$$

is a statement of the result that any quantity can be expressed as a linear combination of its realm matrix and its partition vector. (Since $\boldsymbol{R}(\underline{X}_N)$ contains all the possible values of the quantity $\underline{X}_N$, one and only one of the events $(\underline{X}_N = \underline{x}_{Ni})$ equals 1, and the others equal 0. This unique event is multiplied by $\underline{x}_{Ni}$ in (2.1), so that the entire product necessarily equals the value of $\underline{X}_N$.)

The subjectivist formulation specifies probability as a number (or an interval) that represents an individual's assessment of their own personal knowledge about an event. Such a probability can be defined in terms of betting behaviour.

**Definition 2.1.8** *Let $X$ be any quantity with bounded discrete realm $\mathcal{R}(X)$. Your* ***prevision*** *for $X$, is the number, $P(X)$, such that you would willingly exchange $sX$ for $sP(X)$ and also $sP(X)$ for $sX$, so long as the scale factor, $s$, is a number for which $|s[x - P(X)]| \leq S$ for every $x \in \mathcal{R}(X)$. The number $S > 0$ is called the* ***scale of your maximum stake****. If $X$ is an event then your prevision for $X$ is also called your* ***probability*** *for the event.*

Note that all of the symbols that appear in Definition 2.1.8 represent unitless numbers. A 'loss' or 'gain' from an exchange can be denominated in any units whatsoever — apples, CDs, books — though it is often easiest to think of units of currency. Then, the scale of your maximum stake is a monetary amount that is large enough to be divisible into many noticeably distinct units, but not so large that your acquisition or loss of it would change your life possibilities inordinately.

There is no such notion as a 'true' unknown probability for an event, nor any requirement that different individuals give the same probability to a given event. However, an individual is not entirely free to assert her previsions in any way whatsoever. For instance, it would not make sense for your probability for an event to be a number outside the interval $[0, 1]$, nor for your specified previsions to be self-contradictory in any way. The following property ensures that such inconsistencies do not occur and is a requirement strong enough to induce the equivalent of most classical probability laws.

**Definition 2.1.9** ***Coherency*** *is a requirement that you do not willingly engage transactions that would yield a net loss for you no matter what the value of the quantities happen to be. In other words, you cannot be made a sure loser.*

**Example 2.1.1** *Let $E$ and $F$ be events for which $F = \tilde{E} = 1 - E$, such as the events that it rains this weekend and does not rain this weekend. Suppose you assert $P(E) = 0.6$, $P(F) = 0.9$. Then you are willing to pay $P(E)$ in return for $E$ and pay $P(F)$ in return for $F$. Your net gain would be*

$$\begin{aligned} E - P(E) + F - P(F) &= E + F - P(E) - P(F) \\ &= 1 - 0.6 - 0.9 \\ &= -0.5 \end{aligned}$$

*no matter whether $(E, F) = (1, 0)$ or $(0, 1)$. Your prevision assertions are incoherent.*

**Theorem 2.1.1** *For any vector of quantities, $\underline{X}_N$, and any two vectors of real-valued constants, $\underline{a}_N$ and $\underline{b}_N$, coherent prevision assertion requires that*

$$P(\underline{a}_N{}^T \underline{X}_N + \underline{b}_N) = \underline{a}_N{}^T P(\underline{X}_N) + \underline{b}_N \,.$$

PROOF: See Lad [56]. □

The following definition introduces a very useful way of describing quantities that are considered to be similar.

**Definition 2.1.10** *Suppose $X_1, \ldots, X_N$ are distinct quantities with identical realms. You are said to regard the quantities $X_1, \ldots, X_N$* ***exchangeably*** *if, for any selection of $n \leq N$ of these quantities, your prevision for the product events of the form $(X_1 = y_1)(X_2 = y_2) \ldots (X_n = y_n)$ is constant for every permutation of the numbers $y_1, \ldots, y_n$.*

**Example 2.1.2** *Suppose $X_1, X_2, X_3$ denote exam marks of three students of similar ability. Your judgement to regard $X_1, X_2, X_3$ exchangeably would imply*

$$\begin{aligned} & P(X_1 = 60, X_2 = 55, X_3 = 62) \\ = \; & P(X_1 = 60, X_2 = 62, X_3 = 55) \\ = \; & P(X_1 = 55, X_2 = 60, X_3 = 62) \\ = \; & P(X_1 = 55, X_2 = 62, X_3 = 60 \\ = \; & P(X_1 = 62, X_2 = 60, X_3 = 55) \\ = \; & P(X_1 = 62, X_2 = 55, X_3 = 60)\,. \end{aligned}$$

**Definition 2.1.11** *Let $X$ be any quantity, and $E$ be any event. Your* **conditional prevision** *for $X$ given $E$, denoted $P(X|E)$, is the number such that you are willing to engage any transaction that would yield you a net gain of the amount $s[XE - EP(X|E)]$, as long as $|s[xe - eP(X|E)]| \leq S$ for every pair of numbers $(e, xe) \in \mathcal{R}(E, XE)$. (Again $S$ denotes the scale of your maximum stake.)*

Hence your conditional prevision $P(X|E)$ is defined as the price at which you are indifferent to engaging in the transaction for $X$, contingent on $E$. For if $E = 1$, your net gain from asserting $P(X|E)$ would equal $s[X - P(X|E)]$, whereas if $E = 0$, your net gain would equal 0.

**Definition 2.1.12** *Having specified your conditional prevision, $P(X|E)$, the* **conditional quantity $X$ given $E$,** *denoted by $(X|E)$, is defined as*

$$(X|E) = XE + (1 - E)P(X|E)\,.$$

Hence the conditional quantity $(X|E)$ is an unusual type of quantity in the sense that it is defined in terms of your prevision for it. Note that

$$(X|E) - P(X|E) = XE - EP(X|E)\,,$$

so that

$$\begin{aligned} P(XE - EP(X|E)) &= P((X|E) - P(X|E)) \\ &= P(X|E) - P(X|E) \\ &= 0\,, \end{aligned} \tag{2.2}$$

by Theorem 2.1.1.

**Theorem 2.1.2** *Let $X$ be any quantity and $E$ be any event. Coherency requires that any asserted previsions satisfy*

$$P(XE) = P(X|E)P(E)\,.$$

PROOF: See Lad [56]. □

Having completed a background summary of the language and notation of subjective statistics, we are now ready to reformulate the problem in this context.

## 2.2 Reformulation of the Problem in a Subjectivist Framework

Let $\underline{Y}_{N+1}$ be the $(N+1) \times 1$ vector of quantities whose $i^{\text{th}}$ component represents the outcome of $\underline{Y}^{(i)}$, $i = 1, \ldots, N+1$. (Here we relax the definition of a quantity to include a vector of numbers resulting from the performance of an operational measurement procedure. Hence $\underline{Y}_{N+1}$ can be thought of as a vector whose entries are vectors.) By (2.1),

$$\underline{Y}_{N+1} = \boldsymbol{R}(\underline{Y}_{N+1}).\underline{Q}\,,$$

where $\boldsymbol{R}(\underline{Y}_{N+1})$ is the $(N+1) \times \left({}^{r+K}C_K\right)^{N+1}$ realm matrix for $\underline{Y}_{N+1}$, and $\underline{Q}$ is the $\left({}^{r+K}C_K\right)^{N+1} \times 1$ vector whose components are the constituent events of the partition generated by $\underline{Y}_{N+1}$. Hence we can write

$$P(\underline{Y}_{N+1}) = \boldsymbol{R}(\underline{Y}_{N+1}).\underline{\mathcal{C}}\,, \tag{2.3}$$

where $\underline{\mathcal{C}} = P(\underline{Q})$ is the $\left({}^{r+K}C_K\right)^{N+1} \times 1$ vector of your probability assertions for the constituent events of the partition generated by $\underline{Y}_{N+1}$, if indeed you would assert them at all. To assign a partial ordering to the columns of $\boldsymbol{R}(\underline{Y}_{N+1})$, let $a_t$ and $b_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of components of type $t$ in columns $j$ and $k$, respectively. Let $T$ be the smallest $t \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $a_t \neq b_t$. Column $j$ precedes column $k$ (where it is understood that $j < k$) if and only if $a_T > b_T$. The components of $\underline{Q}$ and $\underline{\mathcal{C}}$ are ordered correspondingly.

The quantities $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N+1)}$, representing observations from similar groups, have identical realms and may well be regarded exchangeably. This means that (2.3) can be rewritten as

$$P(\underline{Y}_{N+1}) = \boldsymbol{R}(\underline{Y}_{N+1}).\mathcal{M}.\underline{q}\,, \tag{2.4}$$

where $\mathcal{M}$ is an $\left({}^{r+K}C_K\right)^{N+1} \times {}^{(N+{}^{r+K}C_K)}C_{(N+1)}$ sparse matrix (description delayed), and $\underline{q}$ is the ${}^{(N+{}^{r+K}C_K)}C_{(N+1)} \times 1$ vector in the $\left({}^{(N+{}^{r+K}C_K)}C_{(N+1)} - 1\right)$- dimensional simplex that would equal your prevision assertions for the constituents of the partition definable in terms of how many $\underline{Y}^{(i)}$, $i = 1, \ldots, N+1$, are of each of the ${}^{r+K}C_K$ possible types. (The number of such constituents is the number of ways of putting $N+1$ things into ${}^{r+K}C_K$ boxes, namely ${}^{(N+{}^{r+K}C_K)}C_{(N+1)}$.) To assign an ordering to the components of $\underline{q}$, let $q_j$ be

$$P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } 1\right) = a_1, \ldots, \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = a_{r+K C_K}\right)$$

and $q_k$ be

$$P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } 1\right) = b_1, \ldots, \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = b_{r+K C_K}\right).$$

Let $T$ be the smallest $t \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $a_t \neq b_t$. Then $q_j$ precedes $q_k$ (where it is understood that $j < k$) if and only if $a_T > b_T$. Again assuming $q_j$ represents

$$P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } 1\right) = a_1, \ldots, \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = a_{r+K C_K}\right),$$

the $j^{\text{th}}$ column of $\mathcal{M}$ contains zeross except for ${}^{N+1}C_{a_1, a_2, \ldots, a_{r+K C_K}}$ components that equal $1 \Big/ {}^{N+1}C_{a_1, a_2, \ldots, a_{r+K C_K}}$. These nonzero components have the same row indices as the indices of the columns of $\boldsymbol{R}(\underline{Y}_{N+1})$ that contain $a_1$ entries of type 1, $a_2$ entries of type 2, ..., $a_{r+K C_K}$ entries of type ${}^{r+K}C_K$. When columns of $\boldsymbol{R}(\underline{Y}_{N+1})$ are not ordered by the rule mentioned earlier they will come in groups and may be ordered in any way whatsoever. The construction of $\mathcal{M}$ and $\underline{q}$ will be the same regardless.

The following example illustrates this notation.

**Example 2.2.1** *Let $N+1 = 3$, $r = 2$ and $K+1 = 3$. Then ${}^{r+K}C_K = {}^{2+2}C_2 = 6$ and the six possible types, in order, are*

$$\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

*The* $3 \times 216$ *matrix* $\boldsymbol{R}(\underline{Y}_3)$, *the* $216 \times 56$ *matrix* $\mathcal{M}$ *and the* $56 \times 1$ *vector* $\underline{q}$, *whose components sum to 1, now appear.*

$$\boldsymbol{R}(\underline{Y}_3) = \begin{bmatrix} \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \cdots & \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \\ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \cdots & \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \\ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} & \cdots & \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \end{bmatrix}$$

$$\mathcal{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1/3 & 0 & \cdots & 0 \\ 0 & 1/3 & 0 & \cdots & 0 \\ 0 & 1/3 & 0 & \cdots & 0 \\ 0 & 0 & 1/3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\underline{q} = \begin{pmatrix} P\left(\underline{Y}^{(1)} = \underline{Y}^{(2)} = \underline{Y}^{(3)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right) \\ P\left(\textit{one of } \underline{Y}^{(i)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \textit{ other two } \underline{Y}^{(i)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right) \\ P\left(\textit{one of } \underline{Y}^{(i)} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \textit{ other two } \underline{Y}^{(i)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right) \\ \vdots \\ P\left(\underline{Y}^{(1)} = \underline{Y}^{(2)} = \underline{Y}^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}\right) \end{pmatrix}$$

*Note that the ordering of columns 1, 2 and 3 of* $\boldsymbol{R}(\underline{Y}_3)$ *has been chosen arbitrarily, but that column 2 of* $\mathcal{M}$ *is not affected by this choice.*

Recall that it is often of great interest to estimate the predictive probabilities

$$P\left(\underline{Y}^{(N+1)} = \underline{y} \,\middle|\, \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right),$$

where $\underline{y}$ is a given type. As (2.4) reflects, the judgement to regard $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N+1)}$ exchangeably means that the information provided to you by the individual outcomes of all of the first $N$ groups can be summarised by recording how many $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, are of each of the ${}^{r+K}C_K$ possible types.

**Definition 2.2.1** *A* ***histogram*** *is a record of how many* $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, *are of each of the* ${}^{r+K}C_K$ *possible types.*

The number of possible histograms is the number of ways of putting $N$ things into ${}^{r+K}C_K$ boxes, namely ${}^{(N+{}^{r+K}C_K-1)}C_N$. It will also be convenient to assign an ordering to the histograms. Suppose $x_t$ and $y_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in histograms $Hu$ and $Hv$, respectively. Let $T$ be the smallest $t \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $x_t \neq y_t$. $Hu$ precedes $Hv$ in order (where it is understood that $u < v$) if and only if $x_T > y_T$. Hence, for example, $H1$ consists entirely of $\underline{Y}^{(i)}$ of type 1.

Let $\underline{Y}_N$ be the $N \times 1$ vector whose $i^{\text{th}}$ component represents the outcome of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, and $H(\underline{Y}_N)$ be its associated histogram. Let

$$p_{t,H} = P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_N) = H)\right), \qquad t = 1, \ldots, {}^{r+K}C_K\,. \tag{2.5}$$

Obviously,

$$\sum_{t=1}^{{}^{r+K}C_K} p_{t,H} = 1, \qquad \forall H\,,$$

so that asserting all of the probabilities in (2.5) is equivalent to asserting the entire predictive probability distribution for $\underline{Y}^{(N+1)}$ given $H$.

Let $\mathcal{H}$ be the set of all histograms for which you are willing, *a priori*, to assert your predictive probabilities as in (2.5). For each histogram $H \in \mathcal{H}$, define ${}^{r+K}C_K$ quantities of the form

$$\left(\underline{Y}^{(N+1)} = \text{type } t\right)(H(\underline{Y}_N) = H) - p_{t,H}(H(\underline{Y}_N) = H), \qquad t = 1, \ldots, {}^{r+K}C_K\,.$$

By (2.2), your prevision for each of these quantities is zero,

$$P\left(\left(\underline{Y}^{(N+1)} = \text{type } t\right)(H(\underline{Y}_N) = H) - p_{t,H}(H(\underline{Y}_N) = H)\right) = 0\,.$$

Suppose $\mathcal{H} = \{H1, H2, \ldots, Hz\}$. Thus define a vector of quantities, $\underline{B}$, where

$$\underline{B} = \begin{pmatrix} \left(\underline{Y}^{(N+1)} = \text{type } 1\right)(H(\underline{Y}_N) = H1) - p_{1,H1}(H(\underline{Y}_N) = H1) \\ \left(\underline{Y}^{(N+1)} = \text{type } 2\right)(H(\underline{Y}_N) = H1) - p_{2,H1}(H(\underline{Y}_N) = H1) \\ \vdots \\ \left(\underline{Y}^{(N+1)} = \text{type } {}^{r+K}C_K\right)(H(\underline{Y}_N) = H1) - p_{{}^{r+K}C_K,H1}(H(\underline{Y}_N) = H1) \\ \left(\underline{Y}^{(N+1)} = \text{type } 1\right)(H(\underline{Y}_N) = H2) - p_{1,H2}(H(\underline{Y}_N) = H2) \\ \vdots \\ \left(\underline{Y}^{(N+1)} = \text{type } {}^{r+K}C_K\right)(H(\underline{Y}_N) = H2) - p_{{}^{r+K}C_K,H2}(H(\underline{Y}_N) = H2) \\ \vdots \\ \left(\underline{Y}^{(N+1)} = \text{type } 1\right)(H(\underline{Y}_N) = Hz) - p_{1,Hz}(H(\underline{Y}_N) = Hz) \\ \vdots \\ \left(\underline{Y}^{(N+1)} = \text{type } {}^{r+K}C_K\right)(H(\underline{Y}_N) = Hz) - p_{{}^{r+K}C_K,Hz}(H(\underline{Y}_N) = Hz) \end{pmatrix}$$

is a $z\left({}^{r+K}C_K\right) \times 1$ vector, satisfying $P(\underline{B}) = \underline{0}$.

Then

$$\underline{B} = R(\underline{B}).\underline{Q}\,,$$

where $R(\underline{B})$ is a $z\left({}^{r+K}C_K\right) \times \left({}^{r+K}C_K\right)^{N+1}$ matrix with entry $[R(\underline{B})]_{cd}$ in row $c$, column $d$ as follows. Consider the $j^{\text{th}}$ row of $R(\underline{B})$. Suppose $j = (h-1)\left({}^{r+K}C_K\right) + m$, where $h \in \{1, \ldots, z\}, m \in \left\{1, \ldots, {}^{r+K}C_K\right\}$. Then $[R(\underline{B})]_{jl} = 0$ unless the histogram of the first $N$ components of column $l$ of $\boldsymbol{R}(\underline{Y}_{N+1})$ is histogram $Hh$, being the $h^{\text{th}}$ in order, from $\mathcal{H}$. In that case, $[R(\underline{B})]_{jl}$ equals the value of the event (the last component of column $l$ of $\boldsymbol{R}(\underline{Y}_{N+1})$ is type $m$) less $p_{m,Hh}$. The possible values of the entries in row $j$ of $R(\underline{B})$ are thus $0, -p_{m,Hh}$ and $1 - p_{m,Hh}$.

Hence,

$$\begin{pmatrix} \underline{Y}_{N+1} \\ \underline{B} \end{pmatrix} = \begin{pmatrix} \boldsymbol{R}(\underline{Y}_{N+1}) \\ R(\underline{B}) \end{pmatrix}.\underline{Q}$$

and

$$P\begin{pmatrix} \underline{Y}_{N+1} \\ \underline{B} \end{pmatrix} = \begin{pmatrix} \boldsymbol{R}(\underline{Y}_{N+1}) \\ R(\underline{B}) \end{pmatrix}.\underline{C}$$

$$= \begin{pmatrix} \boldsymbol{R}(\underline{Y}_{N+1}) \\ R(\underline{B}) \end{pmatrix} .\mathcal{M}.\underline{q}$$

according to the exchangeability assertion. Of most consequence is the resulting equation

$$\begin{aligned} P(\underline{B}) &= R(\underline{B}).\mathcal{M}.\underline{q} \\ &= \underline{0}, \end{aligned} \tag{2.6}$$

which shows that the act of asserting your various predictive probabilities places further restrictions on the components of the vector $\underline{q}$, which must already sum to 1.

In order to understand the exact form of the equations in (2.6), it is necessary to describe further the matrix $R(\underline{B})$, in particular row $j$ where $j = (h-1)\binom{r+K}{K}+m$ with $h, m$ as before. Obviously the first $N$ components of a column of $\boldsymbol{R}(\underline{Y}_{N+1})$ can be histogram $Hh$ only if the histogram of all $N+1$ components is $Hh$ plus an extra type 1, or $Hh$ plus an extra type 2, ..., or $Hh$ plus an extra type ${}^{r+K}C_K$. This implies that there are ${}^{r+K}C_K$ 'groups' of columns containing the nonzero entries in row $j$ of $R(\underline{B})$. Group $t$, $t = 1, \ldots, {}^{r+K}C_K$, will have the same column indices as those columns of $\boldsymbol{R}(\underline{Y}_{N+1})$ whose histogram equals $Hh$ plus an extra type $t$. Entries in group $t$ corresponding to these columns of $\boldsymbol{R}(\underline{Y}_{N+1})$ where the *last* component is *not* type $t$ will be zero (for then the histogram of the first $N$ components is not $Hh$). The other nonzero entries in group $t$ will all equal $1 - p_{m,Hh}$ if $t = m$, and $-p_{m,Hh}$ otherwise. (Note that $(t = m)$ will be true for exactly one $t \in \left\{1, \ldots, {}^{r+K}C_K\right\}$.) The number of nonzero entries in each of these ${}^{r+K}C_K$ groups in row $j$ is therefore the number of distinct orderings of the $N$ types in histogram $Hh$. Letting $x_s$, $s = 1, \ldots, {}^{r+K}C_K$, denote how many type $s$ are in $Hh$, this number is ${}^{N}C_{x_1, x_2, \ldots, x_{r+K C_K}}$.

Consider the $\binom{r+K}{K}^{N+1} \times 1$ vector $\mathcal{M}\underline{q}$. Suppose column $l$ of $\boldsymbol{R}(\underline{Y}_{N+1})$ contains $a_1$ entries of type 1, ..., $a_{r+K C_K}$ entries of type ${}^{r+K}C_K$. Then the $l^{\text{th}}$ component of $\mathcal{M}\underline{q}$ is $\left(1 \Big/ {}^{N+1}C_{a_1, \ldots, a_{r+K C_K}}\right) q_*$ where

$$q_* = P\left(\sum_{i=1}^{N+1} \left(\underline{Y}^{(i)} = \text{type } 1\right) = a_1, \ \ldots, \ \sum_{i=1}^{N+1} \left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = a_{r+K C_K}\right).$$

Now the positive entries in row $j$ of $R(\underline{B})$ occur in ${}^{N}C_{x_1, \ldots, x_{r+K C_K}}$ of the columns having the same indices as those of $\boldsymbol{R}(\underline{Y}_{N+1})$ that form histogram $Hh$ plus an extra type 1, and in ${}^{N}C_{x_1, \ldots, x_{r+K C_K}}$ of the columns having the same indices as those of

$\boldsymbol{R}(\underline{Y}_{N+1})$ that form histogram $Hh$ plus an extra type 2, ..., and in ${}^{N}C_{x_1,\ldots,x_{r+K}C_K}$ of the columns having the same indices as those of $\boldsymbol{R}(\underline{Y}_{N+1})$ that form histogram $Hh$ plus an extra type ${}^{r+K}C_K$. Hence the $j^{\text{th}}$ equation of (2.6) is

$$\sum_{t=1}^{{}^{r+K}C_K}\left[\frac{{}^{N}C_{x_1,\ldots,x_{r+K}C_K}}{{}^{N+1}C_{x_1,\ldots,x_t+1,\ldots,x_{r+K}C_K}}\left(\delta_{mt}-p_{m,Hh}\right)q_{Hh(t)}\right]=0\,,$$

where

$$\delta_{mt}=\begin{cases}1, & t=m\\ 0, & \text{otherwise}\end{cases}$$

and

$$q_{Hh(t)}=P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)}=\text{type }1\right)=x_1,\ldots,\right.$$
$$\left.\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)}=\text{type }t\right)=x_t+1,\ldots,\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)}=\text{type }{}^{r+K}C_K\right)=x_{r+K}C_K\right)\,. \tag{2.7}$$

Simplification gives the $j^{\text{th}}$ equation of (2.6) to be

$$\sum_{t=1}^{{}^{r+K}C_K}\left[\frac{x_t+1}{N+1}\left(\delta_{mt}-p_{m,Hh}\right)q_{Hh(t)}\right]=0$$

or

$$\sum_{t=1}^{{}^{r+K}C_K}\left[f_{t,Hh}\left(\delta_{mt}-p_{m,Hh}\right)q_{Hh(t)}\right]=0\,,$$

where

$$f_{t,Hh}=\frac{x_t+1}{N+1} \tag{2.8}$$

and $x_t$ denotes the number of type $t$ in histogram $Hh$. Note

$$0<f_{t,Hh}\leq 1,\qquad \forall\, t\in\left\{1,\ldots,{}^{r+K}C_K\right\},h\in\{1,\ldots,z\}\,.$$

Note that if the first $N$ components of column $l$ of $\boldsymbol{R}(\underline{Y}_{N+1})$ do not form a histogram from $\mathcal{H}$, then column $l$ of $R(\underline{B})$ contains all zeros. This means that components of $\underline{q}$ that represent your previsions for outcomes of the $N+1$ groups including among them no subset of $N$ that form a histogram from $\mathcal{H}$ do not appear in any of the equations of (2.6).

Recalling that $j = (h-1)\binom{r+K}{K}+m$, with $h \in \{1,\ldots,z\}, m \in \left\{1,\ldots,{}^{r+K}C_K\right\}$, we have deduced the following. Associated with each histogram $H \in \mathcal{H}$ there is a 'block' of ${}^{r+K}C_K$ equations, the $i^{\text{th}}$ equation being of the form

$$\sum_{t=1}^{{}^{r+K}C_K} \left[f_{t,H}\left(\delta_{it} - p_{i,H}\right) q_{H(t)}\right] = 0\,,$$

where

$$\delta_{it} = \begin{cases} 1, & t = i \\ 0, & \text{otherwise} \end{cases}$$

and $p_{t,H}, q_{H(t)}$ and $f_{t,H}$ are as in (2.5), (2.7) and (2.8), respectively.

**Example 2.2.2** *Let $N+1 = 3$, $r = 2$, $K+1 = 3$ and let $H1$ be the histogram formed when $\underline{Y}^{(1)}$ and $\underline{Y}^{(2)}$ are both of type 1 (see Example 2.2.1). The block of six equations associated with $H1$ is*

$$\begin{aligned}
(1-p_{1,H1})q_1 - \frac{1}{3}p_{1,H1}q_2 - \frac{1}{3}p_{1,H1}q_3 - \frac{1}{3}p_{1,H1}q_4 - \frac{1}{3}p_{1,H1}q_5 - \frac{1}{3}p_{1,H1}q_6 &= 0\\
-p_{2,H1}q_1 + \frac{1}{3}(1-p_{2,H1})q_2 - \frac{1}{3}p_{2,H1}q_3 - \frac{1}{3}p_{2,H1}q_4 - \frac{1}{3}p_{2,H1}q_5 - \frac{1}{3}p_{2,H1}q_6 &= 0\\
-p_{3,H1}q_1 - \frac{1}{3}p_{3,H1}q_2 + \frac{1}{3}(1-p_{3,H1})q_3 - \frac{1}{3}p_{3,H1}q_4 - \frac{1}{3}p_{3,H1}q_5 - \frac{1}{3}p_{3,H1}q_6 &= 0\\
-p_{4,H1}q_1 - \frac{1}{3}p_{4,H1}q_2 - \frac{1}{3}p_{4,H1}q_3 + \frac{1}{3}(1-p_{4,H1})q_4 - \frac{1}{3}p_{4,H1}q_5 - \frac{1}{3}p_{4,H1}q_6 &= 0\\
-p_{5,H1}q_1 - \frac{1}{3}p_{5,H1}q_2 - \frac{1}{3}p_{5,H1}q_3 - \frac{1}{3}p_{5,H1}q_4 + \frac{1}{3}(1-p_{5,H1})q_5 - \frac{1}{3}p_{5,H1}q_6 &= 0\\
-p_{6,H1}q_1 - \frac{1}{3}p_{6,H1}q_2 - \frac{1}{3}p_{6,H1}q_3 - \frac{1}{3}p_{6,H1}q_4 - \frac{1}{3}p_{6,H1}q_5 + \frac{1}{3}(1-p_{6,H1})q_6 &= 0\,,
\end{aligned}$$

*where*

$$\begin{aligned}
p_{1,H1} &= P\left(\underline{Y}^{(3)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} \middle| \left(H(\underline{Y}_2) = H1\right)\right)\\
&\vdots\\
p_{6,H1} &= P\left(\underline{Y}^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \middle| \left(H(\underline{Y}_2) = H1\right)\right)
\end{aligned}$$

*and*

$$q_1 = P\left(\underline{Y}^{(1)} = \underline{Y}^{(2)} = \underline{Y}^{(3)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right)$$

$$q_2 = P\left(\textit{one of } \underline{Y}^{(i)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},\ \textit{other two } \underline{Y}^{(i)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right)$$

$$\vdots$$

$$q_6 = P\left(\textit{one of } \underline{Y}^{(i)} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix},\ \textit{other two } \underline{Y}^{(i)} = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}\right).$$

## 2.3 Coherency Induced Equations

The purpose of this section is to learn more about the structure of the system of equations that has just been derived, in order to see what implications coherency has for their solution. We shall find that a feature that conditioning histograms are 'linked' plays a central role in characterising types of solutions.

**Theorem 2.3.1** *The last of the equations in the block associated with a histogram, $H$, is linearly dependent on the others and can always be disregarded. The remaining ${}^{r+K}C_K - 1$ homogeneous equations in ${}^{r+K}C_K$ variables are linearly independent.*

PROOF: Let $A$ be the ${}^{r+K}C_K \times {}^{r+K}C_K$ matrix of coefficients of the equations, $a_{it}$ denoting the entry in row $i$, column $t$. For simplicity we will drop the subscripts on $f$ and $p$ that denote dependence on the histogram, $H$, so that

$$a_{it} = f_t(\delta_{it} - p_i)q_{H(t)}, \qquad i,t \in \left\{1,\ldots,{}^{r+K}C_K\right\}.$$

Now,

$$\begin{aligned} -\sum_{i=1}^{{}^{r+K}C_K-1} a_{it} &= -\sum_{i=1}^{{}^{r+K}C_K-1} \left[f_t(\delta_{it} - p_i)q_{H(t)}\right] \\ &= -f_t\left(\sum_{i=1}^{{}^{r+K}C_K-1} \delta_{it} - \sum_{i=1}^{{}^{r+K}C_K-1} p_i\right) q_{H(t)} \end{aligned}$$

$$\begin{aligned} &= -f_t \left(1 - \delta_{r+{}^KC_K,t} - \sum_{i=1}^{r+{}^KC_K-1} p_i\right) q_{H(t)} \\ &= f_t \left(\delta_{r+{}^KC_K,t} - p_{r+{}^KC_K}\right) q_{H(t)} \\ &= a_{r+{}^KC_K,t}\,. \end{aligned}$$

Thus the last row of $A$ is negative the sum of the other rows.

Let $B$ be the $\left({}^{r+K}C_K - 1\right) \times {}^{r+K}C_K$ submatrix formed by the first ${}^{r+K}C_K - 1$ rows of $A$, and $\underline{d}$ be a $\left({}^{r+K}C_K - 1\right) \times 1$ vector of constants. Suppose $\underline{d}^T B = \underline{0}^T$. Then

$$\begin{aligned} & d_1 f_1(1-p_1) + d_2 f_1(-p_2) + \cdots + d_{r+{}^KC_K-1} f_1(-p_{r+{}^KC_K-1}) \\ = \; & f_1 \left(d_1 - \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) \\ = \; & 0\,. \end{aligned}$$

Similarly,

$$\begin{aligned} f_2 \left(d_2 - \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) &= 0 \\ &\vdots \\ f_{r+{}^KC_K-1} \left(d_{r+{}^KC_K-1} - \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) &= 0\,. \end{aligned}$$

Also,

$$\begin{aligned} & d_1 f_{r+{}^KC_K-1}(-p_1) + d_2 f_{r+{}^KC_K-1}(-p_2) + \cdots + d_{r+{}^KC_K-1} f_{r+{}^KC_K-1}(-p_{r+{}^KC_K-1}) \\ = \; & f_{r+{}^KC_K} \left(- \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) \\ = \; & 0\,. \end{aligned}$$

Since $f_t > 0$, $t = 1, \ldots, {}^{r+K}C_K$,

$$\left(d_1 - \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) = \cdots = \left(d_{r+{}^KC_K-1} - \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) = \left(- \sum_{i=1}^{r+{}^KC_K-1} d_i p_i\right) = 0\,,$$

which implies that $\underline{d} = \underline{0}$. Thus the rows of $B$ are linearly independent. □

**Theorem 2.3.2** *Providing all of the predictive probabilities, $p_{t,H}, t = 1, \ldots, {}^{r+K}C_K$, are strictly positive, the unique form of the solution to the block of equations associated with a histogram, $H$, is*

$$q_{H(t)} = \frac{f_{*,H} p_{t,H}}{f_{t,H} p_{*,H}} q_{H(*)}\,, \qquad t = 1, \ldots, {}^{r+K}C_K\,, \tag{2.9}$$

*where* $q_{H(*)} \in \{q_{H(1)}, \ldots, q_{H({}^{r+K}C_K)}\}$.

PROOF: Let $A$ be the ${}^{r+K}C_K \times {}^{r+K}C_K$ matrix of coefficients of the equations. Standard matrix theory [5] states that

$$\text{nullity}(A) = \text{number of columns of } A - \text{rank}(A)\,.$$

That is, the number of linearly independent solutions to the block of equations is ${}^{r+K}C_K - \left({}^{r+K}C_K - 1\right) = 1$. Suppose (2.9) holds. Once again we will drop the subscripts on $f$ and $p$ that denote dependence on $H$. Then the $i^{\text{th}}$ equation becomes

$$\begin{aligned}
\sum_{t=1}^{{}^{r+K}C_K} \left[ f_t(\delta_{it} - p_i) \frac{f_* p_t}{f_t p_*} q_{H(*)} \right] &= \frac{f_*}{p_*} q_{H(*)} \sum_{t=1}^{{}^{r+K}C_K} [(\delta_{it} - p_i) p_t] \\
&= \frac{f_*}{p_*} q_{H(*)} \left( \sum_{\substack{t=1 \\ t \neq i}}^{{}^{r+K}C_K} -p_i p_t + (1 - p_i) p_i \right) \\
&= \frac{f_*}{p_*} q_{H(*)} p_i \left( 1 - p_i - \sum_{\substack{t=1 \\ t \neq i}}^{{}^{r+K}C_K} p_t \right) \\
&= 0\,,
\end{aligned}$$

for $i = 1, \ldots, {}^{r+K}C_K$, verifying that (2.9) is the desired solution. □

Note that the proof of Theorem 2.3.2 is independent of the choice of $q_{H(*)}$. Thus all of the variables in a given block can be expressed as nonzero multiples of just one of them, and the choice of that one variable is arbitrary.

**Shorthand Notation**: It will be useful to introduce a shorthand notation for the rather lengthy expression in (2.7). Letting $x_s$, $s = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $s$ in histogram $H$,

$$q_{H(t)} = P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } 1\right) = x_1, \ldots, \right. \tag{2.10}$$
$$\left. \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } t\right) = x_t + 1, \ldots, \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = x_{r+{}^{K}C_K}\right)$$

may henceforth be written as

$$P(x_1, \ldots, x_t + 1, \ldots, x_{r+{}^{K}C_K}).$$

**Proposition 2.3.3** *The last two variables in the block of equations associated with a histogram, $H$, are consecutive components of $\underline{q}$.*

PROOF: Let $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$. Let $q_a$ denote the second to last variable in the block of equations and $q_b$ the last variable. Then

$$q_a = q_{H({}^{r+K}C_K - 1)} = P(x_1, \ldots, x_{r+{}^{K}C_K - 2}, x_{r+{}^{K}C_K - 1} + 1, x_{r+{}^{K}C_K})$$

and

$$q_b = q_{H({}^{r+K}C_K)} = P(x_1, \ldots, x_{r+{}^{K}C_K - 2}, x_{r+{}^{K}C_K - 1}, x_{r+{}^{K}C_K} + 1).$$

Suppose there is a variable $q_c$ ordered such that $a \leq c \leq b$. Let $q_c$ represent

$$P(y_1, \ldots, y_{r+{}^{K}C_K - 2}, y_{r+{}^{K}C_K - 1}, y_{r+{}^{K}C_K}).$$

If $x_t \neq y_t$ for some $t \in \left\{1, \ldots, {}^{r+K}C_K - 2\right\}$, then either $c < \min(a, b)$ or $c > \max(a, b)$. Hence $q_c$ can only differ from $q_a$ and $q_b$ in the last two components. For $q_c$ to follow $q_a$ and precede $q_b$ it must be that

$$x_{r+{}^{K}C_K - 1} \leq y_{r+{}^{K}C_K - 1} \leq x_{r+{}^{K}C_K - 1} + 1.$$

Suppose $y_{r+{}^{K}C_K - 1} = x_{r+{}^{K}C_K - 1}$. Then

$$\sum_{t=1}^{r+{}^{K}C_K} x_t = \sum_{t=1}^{r+{}^{K}C_K} y_t = N + 1$$
$$\Rightarrow \quad y_{r+{}^{K}C_K} = x_{r+{}^{K}C_K} + 1$$
$$\Rightarrow \quad q_c = q_b$$
$$\Rightarrow \quad c = b.$$

On the other hand, if $y_{r+KC_K-1} = x_{r+KC_K-1} + 1$, then

$$\begin{aligned} & \sum_{t=1}^{r+KC_K} x_t = \sum_{t=1}^{r+KC_K} y_t = N+1 \\ \Rightarrow \quad & y_{r+KC_K} = x_{r+KC_K} \\ \Rightarrow \quad & q_c = q_a \\ \Rightarrow \quad & c = a\,. \end{aligned}$$

In either case it then follows that $b = a + 1$. □

**Theorem 2.3.4** *The blocks of equations associated with two histograms, $Hu$ and $Hv$, have at most one variable in common.*

PROOF: Let $x_t$ and $y_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $Hu$ and $Hv$, respectively. Then the ${}^{r+K}C_K$ variables in the equations associated with $Hu$ represent

$$P(x_1+1, x_2, \ldots, x_{r+KC_K}),\ P(x_1, x_2+1, \ldots, x_{r+KC_K}),\ \ldots,\ P(x_1, x_2, \ldots, x_{r+KC_K}+1)$$

and the ${}^{r+K}C_K$ variables in the equations associated with $Hv$ represent

$$P(y_1+1, y_2, \ldots, y_{r+KC_K}),\ P(y_1, y_2+1, \ldots, y_{r+KC_K}),\ \ldots,\ P(y_1, y_2, \ldots, y_{r+KC_K}+1)\,.$$

Note that

$$\sum_{t=1}^{r+KC_K} x_t = \sum_{t=1}^{r+KC_K} y_t = N$$

implies that $Hu$ and $Hv$ cannot differ in just one component. If $Hu$ and $Hv$ differ in three or more components then it is obvious that none of the variables associated with $Hu$ is also associated with $Hv$. Suppose $Hu$ and $Hv$ are such that

$$\begin{aligned} y_i &= x_i - c_1 \\ y_j &= x_j - c_2 \\ y_t &= x_t, \qquad t \neq i \text{ or } j,\ t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, \end{aligned}$$

where $1 \leq i < j \leq {}^{r+K}C_K$ and $|c_1|, |c_2| \in \{1, \ldots, N\}$. Then

$$\begin{aligned} N &= \sum_{t=1}^{r+K C_K} y_t \\ &= \sum_{t=1}^{r+K C_K} x_t - c_1 - c_2 \\ &= N - c_1 - c_2 \\ \Rightarrow c_2 &= -c_1 . \end{aligned}$$

Assuming, without loss of generality, that $Hu$ is ordered before $Hv$, then $c_1 > 0$. If $2 \leq c_1 \leq N$, then $y_i + 1 = x_i - c_1 + 1 < x_i$ so there cannot be a variable common to $Hu$ and $Hv$. Suppose $c_1 = 1$. Then there is exactly one shared variable, namely the $j^{\text{th}}$ of those associated with $Hu$ and the $i^{\text{th}}$ of those associated with $Hv$. The shared variable is

$$q_{Hu(j)} = q_{Hv(i)} = P(x_1, \ldots, x_i, \ldots, x_j + 1, \ldots, x_{r+K C_K}) .$$

□

**Definition 2.3.1** *Two histograms are said to be **linked** if their equation blocks have a variable in common. The histograms may then be referred to as **adjacent** or **neighbouring** histograms.*

It is worth reviewing, from the proof of Theorem 2.3.4, the structure that is required for two histograms, $Hu$ and $Hv$, to be linked. Letting $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $Hu$, being the first in order of the two histograms, they may be written

$$\begin{aligned} x_1, \ldots, x_i, \ldots, x_j, \ldots, x_{r+K C_K} \qquad & Hu \\ x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_{r+K C_K} \qquad & Hv , \end{aligned}$$

where $1 \leq i < j \leq {}^{r+K}C_K$, assuming all of their components are nonnegative.

**Definition 2.3.2** *A system made up of the blocks of equations associated with histograms from some set, $\mathcal{H}$, is said to be **linked** if any two histograms in $\mathcal{H}$ are either linked directly, or indirectly through other linked histograms from $\mathcal{H}$. Alternatively, the set $\mathcal{H}$ is said to be **linked**.*

This idea of a linked set of histograms will turn out to be important when considering the solution to a system of equations, but it also seems natural that such a set (or sets) may be produced when you specify those histograms for which you are willing to assert your predictive probabilities in (2.5). You may feel willing to assert these probabilities when the first $N$ groups produce a histogram in some 'range', 'ball-park' or 'neighbourhood' with which you feel comfortable. Such histograms would be likely to be similar to each other and therefore linked either directly or indirectly.

**Theorem 2.3.5** *If the system made up of the blocks of equations associated with histograms from some set, $\mathcal{H}$, is linked and all of your asserted predictive probabilities that make up these equations are strictly positive, then either*
*(a) this system has only the trivial solution, or*
*(b) this system has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Let $B$ now denote the number of histograms in $\mathcal{H}$, *i.e.*, the number of blocks of equations in the system. The proof will be by induction on $B$.

Theorem 2.3.2 shows that the theorem holds when $B = 1$.

Suppose $B = 2$ and $\mathcal{H} = \{H1, H2\}$. As the system is linked, Theorem 2.3.4 shows that $H1$ and $H2$ have exactly one variable in common, say $q_a$. By Theorem 2.3.2, all of the other variables in blocks 1 and 2 can be expressed as nonzero multiples of $q_a$, and hence of any one of them. The theorem holds when $B = 2$.

Suppose $B = 3$ and $\mathcal{H} = \{H1, H2, H3\}$. There are three cases to consider.
Case (a) $H1$, $H2$ and $H3$ are all linked through the same variable, $q_a$. Theorem 2.3.4 shows that there can be no other shared variables. By Theorem 2.3.2, all of the other variables in blocks 1, 2 and 3 can be expressed as nonzero multiples of $q_a$, and hence of any one of them.
Case (b) $H1$ and $H2$ are linked through variable $q_a$. $H1$ and $H3$ are linked through variable $q_b$. $H2$ and $H3$ are only linked indirectly, through $H1$. By Theorem 2.3.2, all of the variables in blocks 1 and 2 can be expressed in terms of $q_a$. Similarly, all of the variables in block 3 can be expressed in terms of $q_b$. However, $q_b$ appears in block 1 and so has an expression in terms of $q_a$. Hence all of the variables in blocks 1, 2 and 3 can be expressed as nonzero multiples of $q_a$, and hence of any of them.

Case (c) $H1$ and $H2$ are linked through variable $q_a$. $H1$ and $H3$ are linked through variable $q_b$. $H2$ and $H3$ are also linked directly, through variable $q_c$. By Theorem 2.3.2, all of the variables in blocks 1 and 2, including $q_b$, can be directly expressed in terms of $q_a$. Similarly, all of the variables in block 3, can be expressed in terms of $q_c$. However, $q_c$ appears in block 2 and so has an expression in terms of $q_a$. Hence all of the variables in block 3, including $q_b$, can be indirectly expressed in terms of $q_a$. Either the two representations of $q_b$ in terms of $q_a$ are the same and all of the variables in blocks 1, 2 and 3 can be expressed as nonzero multiples of $q_a$, and hence of any one of them, or there is an inconsistency forcing $q_a$, and hence all of the variables, to be zero. The theorem holds when $B = 3$.

Suppose that the theorem holds when $B \leq b$. Consider a linked system of $b+1$ blocks of equations. Remove one of these blocks, relating to histogram $H$, to form a system of $b$ blocks of equations. This smaller system will be made up of $l \geq 1$ linked subsystems. Take any one of these $l$ subsystems. Now $H$ must have a variable in common with this subsystem, say $q_a$. By Theorem 2.3.2, all of the variables associated with $H$ can be directly expressed in terms of $q_a$. By the inductive assumption, one of the following two possibilities is true. In one case, all of the variables in this subsystem must be zero. Then $q_a = 0$ and those variables in block $H$ are also zero. In the other case, all of the variables in the subsystem can be expressed as nonzero multiples of $q_a$. If any variable other than $q_a$, say $q_b$, appears in both the subsystem and block $H$, then it has two potentially different representations in terms of $q_a$. If they are different, $q_a$, and hence all of the variables in the subsystem and block $H$ must be zero. If not, all of the variables in the subsystem and block $H$ can be written as nonzero multiples of $q_a$. If any of the $l$ subsystems when considered together with block $H$ has only the trivial solution, then so must the other subsystems (by expressing their free variable (if there is one) in terms of a variable from block $H$). Otherwise, each of the $l$ subsystems has a nonzero representation in terms of a distinct variable from block $H$. Each of these distinct variables can be expressed in terms of $q_a$ and hence all of the variables from the whole system of $b+1$ blocks can be expressed as nonzero multiples of any one of them.

The theorem holds for $B = b+1$, if it holds for $B \leq b$. Since it holds for $B = 1$, 2 and 3, it is true for all $B \in \mathbb{N}$ by induction. □

Obviously the question arises of what happens to Theorem 2.3.5 when the set $\mathcal{H}$ is not linked. The important point to note is that, for reasons already mentioned, $\mathcal{H}$ is then likely to be made up of a small number of linked subsets, to each of which Theorem 2.3.5 does apply.

Practically speaking, what does it mean for the system described in Theorem 2.3.5 to have only the trivial solution? All of the components of $\underline{q}$ that appear with nonzero coefficient in the equations must equal zero. This says that due to your judgement to regard $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N+1)}$ exchangeably and the way in which you have asserted your predictive probabilities in (2.5) for all $H \in \mathcal{H}$, coherency requires that your

$$q_{H(t)} = P\left(\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } 1\right) = x_1, \ldots, \right.$$
$$\left.\sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } t\right) = x_t + 1, \ldots, \sum_{i=1}^{N+1}\left(\underline{Y}^{(i)} = \text{type } {}^{r+K}C_K\right) = x_{r+K C_K}\right) = 0,$$

$\forall\, t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, H \in \mathcal{H}$, where $x_s$, $s = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $s$ in $H$. In other words you are required to give zero probability to the outcome of $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N+1)}$ constituting any histogram from $\mathcal{H}$ plus any other type. So in fact, by Theorem 2.1.2, you must give zero probability to $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$ constituting any histogram from $\mathcal{H}$. But these are exactly the situations in which your predictive probabilities are meant to apply! If you wanted to allow each histogram in $\mathcal{H}$ even a tiny positive probability of occurring, your specified predictive probabilities would be incoherent. The logic of coherent conditional probabilities when conditioning on events that are themselves assessed with probability zero involves details that shall not be dealt with here. Nonetheless, asserted conditional probabilities that allow only the trivial solution to the system described in Theorem 2.3.5 should be recognised to force their proponents into this situation. There is nothing incoherent in this in itself, but the irony is that if you assert $P(H(\underline{Y}_N) = H) = 0$, $\forall H \in \mathcal{H}$, then coherency allows you to assert as values for $P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, H(\underline{Y}_N) = H\right)$, $t = 1, \ldots, {}^{r+K}C_K$, *any* non-negative numbers you like, so long as they sum to 1. Obviously if $\mathcal{H}$ is the set of all possible conditioning histograms then your assertions are truly incoherent, for all of the components of $\underline{q}$ appear in the system of equations and there is no solution to the resulting conditions

$$\begin{aligned} \underline{q} &= \underline{0} \\ \underline{q}^T\underline{1} &= 1\,. \end{aligned}$$

It is therefore of great importance to be able to determine whether (a) or (b) of Theorem 2.3.5 is true for a given linked system of equations. The key to answering this question has its origins in the proof of Theorem 2.3.5 and will be developed in the next section.

## 2.4 $n$-Cycles

We begin by studying the smallest system of linked equations that may have only the trivial solution. As the proof of Theorem 2.3.5 shows, this is when the system is made up of three blocks of equations, associated with, say, $H1$, $H2$ and $H3$, in such a way that $H1$ and $H2$, $H1$ and $H3$, and $H2$ and $H3$ are all linked through different variables. Let $H1$ be written

$$x_1, x_2, \ldots, x_{r+K}C_K\,,$$

where $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H1$. Then $H2$ is

$$x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_{r+K}C_K\,,$$

where $1 \leq i < j \leq {}^{r+K}C_K$. In order to find $H3$, consider all of the possible histograms linked to $H1$. They only differ from $H1$ in two components. Suppose these components are $i$ and $j$. Then the possible histograms are

$$\begin{aligned} &x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_{r+K}C_K \qquad h2 \\ &x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_{r+K}C_K \qquad h3\,. \end{aligned}$$

Suppose the components are $i$ and $k, k \neq j$. Without loss of generality assume $j < k$. Then the possible histograms are

$$\begin{aligned} &x_1, \ldots, x_i - 1, \ldots, x_j, \ldots, x_k + 1, \ldots, x_{r+K}C_K \qquad h4 \\ &x_1, \ldots, x_i + 1, \ldots, x_j, \ldots, x_k - 1, \ldots, x_{r+K}C_K \qquad h5\,. \end{aligned}$$

Suppose the components are $j$ and $k, k \neq i$. Without loss of generality assume $j < k$. Then the possible histograms are

$$\begin{aligned} &x_1, \ldots, x_i, \ldots, x_j - 1, \ldots, x_k + 1, \ldots, x_{r+K}C_K \qquad h6 \\ &x_1, \ldots, x_i, \ldots, x_j + 1, \ldots, x_k - 1, \ldots, x_{r+K}C_K \qquad h7\,. \end{aligned}$$

Suppose the components are $l$ and $k$, $l, k \neq i$ or $j$. Without loss of generality assume $j < k < l$. Then the possible histograms are

$$\begin{aligned} &x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k - 1, \ldots, x_l + 1, \ldots, x_{r+KC_K} && h8 \\ &x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k + 1, \ldots, x_l - 1, \ldots, x_{r+KC_K} && h9\,. \end{aligned}$$

Of course the histograms $h2, \ldots, h9$ are only valid providing all of their components are nonnegative. Note that $h2$ is $H2$. Obviously none of $h3$, $h5$, $h6$ , $h8$, $h9$ is linked to $H2$, while $h7$ is linked to $H1$ and $H2$ through the same variable, namely $P(x_1, \ldots, x_i, \ldots, x_j + 1, \ldots, x_k, \ldots, x_{r+KC_K})$. So, by a process of elimination, $h4$ must be $H3$. Thus we have

$$\begin{aligned} &x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k, \ldots, x_{r+KC_K} && H1 \\ &x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_k, \ldots, x_{r+KC_K} && H2 \\ &x_1, \ldots, x_i - 1, \ldots, x_j, \ldots, x_k + 1, \ldots, x_{r+KC_K} && H3\,, \end{aligned} \tag{2.11}$$

where $1 \leq i < j < k \leq {}^{r+K}C_K$. The relative positioning of the indices $i$, $j$, $k$ is necessary to make $H1$ ordered before $H2$ before $H3$. The $j^{\text{th}}$ variable associated with $H1$ is the same as the $i^{\text{th}}$ variable associated with $H2$. The $k^{\text{th}}$ variable associated with $H1$ is the same as the $i^{\text{th}}$ variable associated with $H3$. The $k^{\text{th}}$ variable associated with $H2$ is the same as the $j^{\text{th}}$ variable associated with $H3$. Let

$$\begin{aligned} q_a &= q_{H1(j)} = q_{H2(i)} \\ q_b &= q_{H1(k)} = q_{H3(i)} \\ q_c &= q_{H2(k)} = q_{H3(j)}\,. \end{aligned}$$

We proceed under the assumption that the predictive probabilities conditioning on $H1$, $H2$ and $H3$ are all positive. Then

$$\begin{aligned} q_b &= q_{H1(k)} \\ &= \frac{f_{j,H1}p_{k,H1}}{f_{k,H1}p_{j,H1}} q_{H1(j)} \\ &= \frac{f_{j,H1}p_{k,H1}}{f_{k,H1}p_{j,H1}} q_a\,, \end{aligned}$$

by Theorem 2.3.2. Also,

$$\begin{aligned} q_b &= q_{H3(i)} \\ &= \frac{f_{j,H3}p_{i,H3}}{f_{i,H3}p_{j,H3}} q_{H3(j)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{f_{j,H3}p_{i,H3}}{f_{i,H3}p_{j,H3}} q_{H2(k)} \\
&= \frac{f_{j,H3}p_{i,H3}}{f_{i,H3}p_{j,H3}} \frac{f_{i,H2}p_{k,H2}}{f_{k,H2}p_{i,H2}} q_{H2(i)} \\
&= \frac{f_{j,H3}p_{i,H3}}{f_{i,H3}p_{j,H3}} \frac{f_{i,H2}p_{k,H2}}{f_{k,H2}p_{i,H2}} q_a \,,
\end{aligned}
$$

by Theorem 2.3.2. Hence $q_a = 0$, forcing a trivial solution, unless

$$
\frac{f_{j,H1}p_{k,H1}}{f_{k,H1}p_{j,H1}} = \frac{f_{j,H3}p_{i,H3}}{f_{i,H3}p_{j,H3}} \frac{f_{i,H2}p_{k,H2}}{f_{k,H2}p_{i,H2}} \,,
$$

which can be rewritten as

$$
\frac{p_{k,H1}p_{j,H3}p_{i,H2}}{p_{j,H1}p_{i,H3}p_{k,H2}} = \frac{f_{k,H1}f_{j,H3}f_{i,H2}}{f_{j,H1}f_{i,H3}f_{k,H2}} \,. \tag{2.12}
$$

Consider the right-hand side of (2.12). Using (2.8),

$$
\begin{aligned}
\frac{f_{k,H1}f_{j,H3}f_{i,H2}}{f_{j,H1}f_{i,H3}f_{k,H2}} &= \frac{\left(\frac{x_k+1}{N+1}\right)\left(\frac{x_j+1}{N+1}\right)\left(\frac{x_i}{N+1}\right)}{\left(\frac{x_j+1}{N+1}\right)\left(\frac{x_i}{N+1}\right)\left(\frac{x_k+1}{N+1}\right)} \\
&= 1 \,.
\end{aligned}
$$

Hence the condition for a nontrivial solution to the system of equations associated with $H1$, $H2$ and $H3$ to exist is

$$
\frac{p_{k,H1}p_{j,H3}p_{i,H2}}{p_{j,H1}p_{i,H3}p_{k,H2}} = 1 \,. \tag{2.13}
$$

So we have discovered a necessary condition for any linked system of equations to have a nontrivial solution: For any three histograms from the set generating the equations that are related as in (2.11), your predictive probabilities must satisfy (2.13). Is this also a sufficient condition? Or are there similar restrictions that apply to four, five or more histograms related to each other in some special way? As we shall see, the answers to these questions are no and yes, respectively.

**Definition 2.4.1** *An* ***n-cycle*** *is a linked set of $n \geq 3$ histograms such that each histogram in the set is linked to exactly two of the others.*

The term $n$-cycle is used due to the analogy of the situation here with standard graph theory. The histograms can be thought of as the vertices of a graph in which two vertices are adjacent if and only if their histograms are, *i.e.*, if and only if the

equation blocks associated with the two histograms have a variable in common. Strictly speaking then, an $n$-cycle as described in Definition 2.4.1 corresponds to the vertices of a nonchordal cycle (a cycle that contains no sub-cycles) of length $n$ in the graph. Note also that a set of histograms is linked in our language if the histograms form a connected graph.

We have already encountered one example of an $n$-cycle. The histograms described in (2.11) form a 3-cycle. In fact, as is obvious from the process by which it was constructed, (2.11) represents the only structure that is possible for a 3-cycle.

**Example 2.4.1** *Let $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in histogram $H1$. The following four histograms form a 4-cycle*

$$\begin{aligned} x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k, \ldots, x_l, \ldots, x_{r+K C_K} \qquad & H1 \\ x_1, \ldots, x_i, \ldots, x_j - 1, \ldots, x_k + 1, \ldots, x_l, \ldots, x_{r+K C_K} \qquad & H2 \\ x_1, \ldots, x_i - 1, \ldots, x_j, \ldots, x_k, \ldots, x_l + 1, \ldots, x_{r+K C_K} \qquad & H3 \\ x_1, \ldots, x_i - 1, \ldots, x_j - 1, \ldots, x_k + 1, \ldots, x_l + 1, \ldots, x_{r+K C_K} \qquad & H4, \end{aligned}$$

*where $1 \leq i < j < k < l \leq {}^{r+K}C_K$, assuming all of their components are nonnegative.*

**Diagrammatic Representations**: At times the understanding of a concept or an explanation may be aided by a visual representation of the situation. As an example of how an $n$-cycle may be represented, the 4-cycle in Example 2.4.1 may be drawn as in Figure 2.1 or Figure 2.2.

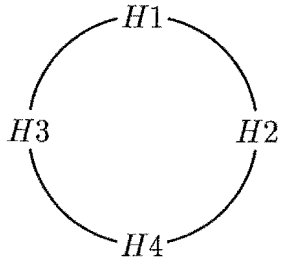



Figure 2.1: Circle Representation of a 4-Cycle

In Figure 2.2 horizontal lines represent the blocks of equations associated with the histograms and vertical lines represent the 'links' or shared variables between histograms. The order in which the vertical lines are drawn corresponds to the relative

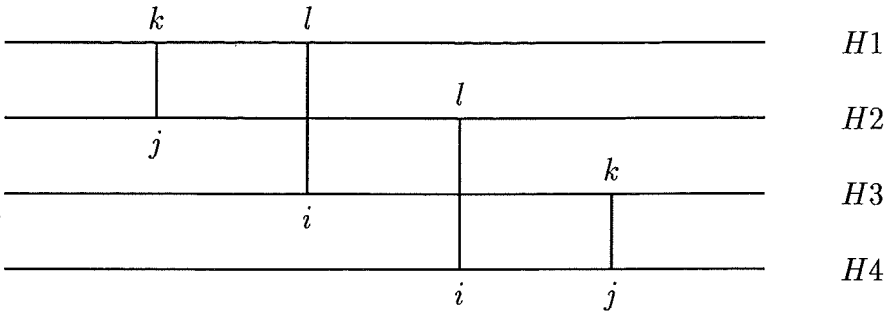



Figure 2.2: Line Representation of a 4-Cycle

ordering of the components of $\underline{q}$ that they represent. The indices above or below a horizontal line at the end of a vertical line indicate which, in order, of the ${}^{r+K}C_K$ variables associated with that histogram is involved in that link.

This idea of considering histograms in the form of a cycle suggests a slightly more straightforward way of discovering whether or not the system of equations that they form has a nontrivial solution. To demonstrate, first consider again the 4-cycle described in Example 2.4.1. Let $q_a = q_{H1(k)} = q_{H2(j)}$. We proceed under the assumption that the predictive probabilities conditioning on $H1$, $H2$, $H3$ and $H4$ are all positive. Then

$$
\begin{aligned}
q_a &= q_{H1(k)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} q_{H1(l)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} q_{H3(i)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} q_{H3(k)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} q_{H4(j)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} \frac{f_{i,H4}p_{j,H4}}{f_{j,H4}p_{i,H4}} q_{H4(i)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} \frac{f_{i,H4}p_{j,H4}}{f_{j,H4}p_{i,H4}} q_{H2(l)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} \frac{f_{i,H4}p_{j,H4}}{f_{j,H4}p_{i,H4}} \frac{f_{j,H2}p_{l,H2}}{f_{l,H2}p_{j,H2}} q_{H2(j)} \\
&= \frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}} \frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}} \frac{f_{i,H4}p_{j,H4}}{f_{j,H4}p_{i,H4}} \frac{f_{j,H2}p_{l,H2}}{f_{l,H2}p_{j,H2}} q_a ,
\end{aligned}
$$

by Theorem 2.3.2. Hence $q_a = 0$, forcing a trivial solution, unless

$$\frac{f_{l,H1}p_{k,H1}}{f_{k,H1}p_{l,H1}}\frac{f_{k,H3}p_{i,H3}}{f_{i,H3}p_{k,H3}}\frac{f_{i,H4}p_{j,H4}}{f_{j,H4}p_{i,H4}}\frac{f_{j,H2}p_{l,H2}}{f_{l,H2}p_{j,H2}} = 1\,.$$

However,

$$\begin{aligned}\frac{f_{l,H1}f_{k,H3}f_{i,H4}f_{j,H2}}{f_{k,H1}f_{i,H3}f_{j,H4}f_{l,H2}} &= \frac{\left(\frac{x_l+1}{N+1}\right)\left(\frac{x_k+1}{N+1}\right)\left(\frac{x_i}{N+1}\right)\left(\frac{x_j}{N+1}\right)}{\left(\frac{x_k+1}{N+1}\right)\left(\frac{x_i}{N+1}\right)\left(\frac{x_j}{N+1}\right)\left(\frac{x_l+1}{N+1}\right)}\\ &= 1\,.\end{aligned}$$

Hence the condition for a nontrivial solution to the system of equations associated with the 4-cycle in Example 2.4.1 to exist is

$$\frac{p_{k,H1}p_{i,H3}p_{j,H4}p_{l,H2}}{p_{l,H1}p_{k,H3}p_{i,H4}p_{j,H2}} = 1\,. \tag{2.14}$$

The structure of a 3-cycle must always be as in (2.11), however there is no such unique structure for $n$-cycles where $n \geq 4$. For example, letting $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in histogram $H1$, the following four histograms form a 4-cycle distinct in structure to that in Example 2.4.1,

$$\begin{aligned} x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k, \ldots, x_l, \ldots, x_{r+K}C_K & \qquad H1\\ x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k - 1, \ldots, x_l + 1, \ldots, x_{r+K}C_K & \qquad H2\\ x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_k, \ldots, x_l, \ldots, x_{r+K}C_K & \qquad H3\\ x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_k - 1, \ldots, x_l + 1, \ldots, x_{r+K}C_K & \qquad H4\,, \end{aligned}$$

where $1 \leq i < j < k < l \leq {}^{r+K}C_K$, assuming all of their components are nonnegative. This 4-cycle may be drawn as in Figure 2.3. Hence there is no guarantee that condition (2.14) is appropriate for all 4-cycles.

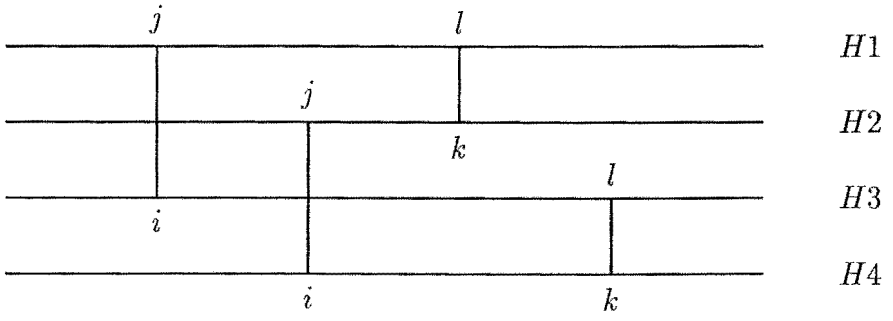



Figure 2.3: Line Representation of Another 4-Cycle

In order to generalise the discoveries made so far, we first note two important points about the nature of any $n$-cycle.

- In going from one histogram to an adjacent histogram changes (of $+1$ and $-1$) are made to exactly two of the $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, components. Hence $2n$ changes are made altogether in going from some arbitrary starting histogram, around the $n$-cycle, until the initial histogram is reached once more.

- In order to 'complete the cycle' and link back to some arbitrary starting histogram, every type that is involved in these changes must be removed as many times as it is added and vice versa.

One obvious consequence is that an $n$-cycle can involve changes to at most $n$ types or components otherwise $2n$ changes would not be enough to reverse all of the additions and subtractions. The following example demonstrates that it is indeed possible for $n$-cycles to exist that involve changes to less than $n$ types.

**Example 2.4.2** *Let $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of type $t$ in histogram $H1$. The following six histograms form a 6-cycle,*

$$\begin{aligned}
&x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k, \ldots, x_l, \ldots, x_m, \ldots, x_{r+K C_K} && H1\\
&x_1, \ldots, x_i, \ldots, x_j, \ldots, x_k, \ldots, x_l - 1, \ldots, x_m + 1, \ldots, x_{r+K C_K} && H2\\
&x_1, \ldots, x_i - 1, \ldots, x_j + 1, \ldots, x_k, \ldots, x_l, \ldots, x_m, \ldots, x_{r+K C_K} && H3\\
&x_1, \ldots, x_i - 1, \ldots, x_j, \ldots, x_k + 1, \ldots, x_l - 1, \ldots, x_m + 1, \ldots, x_{r+K C_K} && H4\\
&x_1, \ldots, x_i - 2, \ldots, x_j + 1, \ldots, x_k + 1, \ldots, x_l, \ldots, x_m, \ldots, x_{r+K C_K} && H5\\
&x_1, \ldots, x_i - 2, \ldots, x_j + 1, \ldots, x_k + 1, \ldots, x_l - 1, \ldots, x_m + 1, \ldots, x_{r+K C_K} && H6,
\end{aligned}$$

*where $1 \leq i < j < k < l < m \leq {}^{r+K}C_K$, assuming all of their components are nonnegative. This 6-cycle may be drawn as in Figure 2.4 or Figure 2.5.*

Let $H1, \ldots, Hn$ be the $n$ histograms, in order, that form a given $n$-cycle. Let $q_a$ be the first, in order, of the two variables associated with $H1$ that are also associated with a neighbouring histogram. Use Theorem 2.3.2 to express $q_a$ in terms of this other variable, say $q_b$. Now $q_b$ is associated with some other histogram, $Hu$, where $u \in \{2, \ldots, n\}$. Express $q_b$ in terms of the other variable associated with $Hu$ that is associated with a neighbouring histogram. Continue in this fashion around the

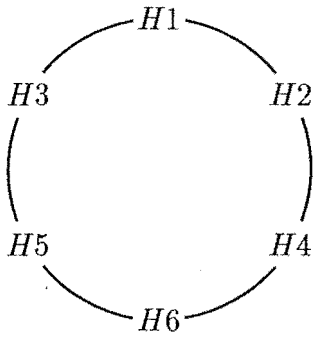



*Figure 2.4: Circle Representation of a 6-Cycle*

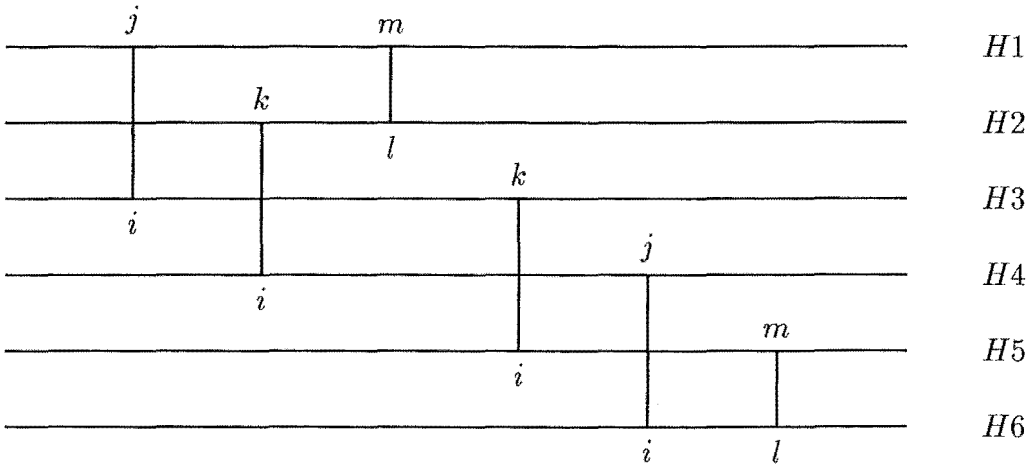



*Figure 2.5: Line Representation of a 6-Cycle*

$n$-cycle until you end up by expressing some variable in terms of $q_a$. Then, due to the nature of an $n$-cycle, the expression created is of the form

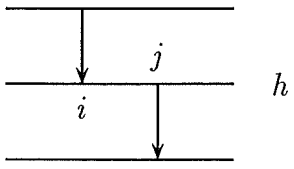
$$q_a = \square \ldots \square q_a \,,$$

where there are $n$ $\square$'s (corresponding one to each histogram) and each $\square$ represents

$$\frac{f_{j,h}p_{i,h}}{f_{i,h}p_{j,h}}, \qquad i \neq j,\ i,j \in \left\{1,\ldots,{}^{r+K}C_K\right\},\ h \in \{H1,\ldots,Hn\}\,. \tag{2.15}$$

This comes from the fact that histogram $h$ shares its $i^{\text{th}}$ variable with the histogram 'preceding' it in the $n$-cycle and its $j^{\text{th}}$ variable with the histogram 'following' it. Or, in other words, to convert the preceding histogram into $h$ a type $i$ is removed and some other type added, and to convert $h$ into the following histogram a type $j$ is added and some other type removed. Consider the overall fraction formed by the product of the $n$ $\square$'s. Suppose there exists $f_{t,h} = (X+1)/(N+1)$ in the numerator. That is, histogram $h$ has $X$ of type $t$ and the histogram following it in the $n$-cycle has $X+1$ of type $t$. Regardless of whether or not further type $t$'s are added (and then removed), to complete the cycle and get back to histogram $h$ there must eventually be a histogram from which removing a type $t$ produces a histogram, say $\mathfrak{h}$, with $X$ of type $t$. That is, there exists $f_{t,\mathfrak{h}} = (X+1)/(N+1)$ in the denominator. Hence all of the $f$ terms cancel. The ratio of the remaining predictive probabilities must equal 1 to allow the existence of a nontrivial solution to the system of equations associated with the given $n$-cycle.

We have discovered a necessary and sufficient condition for a general $n$-cycle to have a nontrivial solution: Each occurrence




in the $n$-cycle contributes $p_{i,h}/p_{j,h}$ to a ratio of predictive probabilities that must equal 1. The arrows on the links merely reflect that the direction taken in going from one histogram to the 'next' in the $n$-cycle determines which of the probabilities goes in the numerator. Of course, had the opposite direction been chosen, the overall ratio would be inverted, producing the same condition when equated to 1.

**Definition 2.4.2** *An n-cycle is said to be* ***satisfied*** *if your asserted predictive probabilities make its required ratio of probabilities equal to* 1.

**Theorem 2.4.1** *Suppose the system made up of the blocks of equations associated with histograms from some set, $\mathcal{H}$, is linked and all of your asserted predictive probabilities that make up these equations are strictly positive. Then this system has a solution in which all of the variables may be expressed as nonzero multiples of any one of them if and only if all of the n-cycles that are formed by histograms from $\mathcal{H}$ are satisfied.*

PROOF: When the cardinality of $\mathcal{H}$ is less than three there are no $n$-cycles to speak of. The proof of Theorem 2.3.5 shows the statement of the theorem to be trivially true in this case. Assume that the cardinality of $\mathcal{H}$ is at least three. Take a histogram $Hu \in \mathcal{H}$ and use Theorem 2.3.2 to express all of the variables associated with $Hu$ in terms of just one of them, say $q_a$. There is at least one histogram, $Hv \in \mathcal{H}$, linked to $Hu$. Use Theorem 2.3.2 to express all of the variables associated with $Hv$ in terms of the variable that is also associated with $Hu$, and hence in terms of $q_a$. Continue adding histograms one at a time in this fashion and using the links to express all of the variables at any given stage in terms of $q_a$. The only way that $q_a$ may be forced to equal zero is if at some point the histogram being added, say $Hw$, has more than one variable in common with the variables already encountered. For then, if $q_b$ and $q_c$ are such variables, $q_b$ can be expressed in terms of $q_c$ by using their representations in terms of $q_a$ that have already been derived and then $q_c$ can be expressed in terms of $q_b$ by applying Theorem 2.3.2 to $Hw$. Hence $q_b$ can be expressed as a nonzero multiple of itself. If the ratio of predictive probabilities that forms this multiple is not equal to 1 then $q_b$, and hence $q_a$, must equal zero. Theorem 2.3.4 implies that $Hw$ is linked to two distinct histograms from the set of histograms already encountered, $\mathcal{H}^*$, which is itself linked. In other words, either $Hw$ and some subset of histograms from $\mathcal{H}^*$ form an $n$-cycle or they form a 'quasi-$n$-cycle' — similar in structure to an $n$-cycle except that each histogram may be linked to more than two others so that links 'across' the $n$-cycle are allowed. If they form an $n$-cycle a nontrivial solution is guaranteed if that $n$-cycle is satisfied. Consider a quasi-$n$-cycle as in Figure 2.6 where $h$ and $\mathfrak{h}$ are the only histograms linked across the $n$-cycle and suppose that $q_{h(i)} = q_{\mathfrak{h}(j)}$. Then effectively this quasi-$n$-cycle is composed of

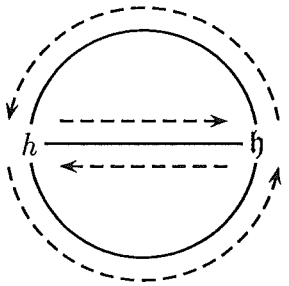



Figure 2.6: A Quasi-$n$-Cyle Representation

an $m$-cycle and an $(n-m+2)$-cycle, where $3 \leq m < n$. Suppose the $m$-cycle and $(n-m+2)$-cycle are satisfied. Then the quasi-$n$-cycle probability ratio is

$$\frac{p_{i,h}\, p_{j,\bar{h}}}{p_{j,\bar{h}}\, p_{i,h}} = 1\,.$$

Hence, if $Hw$ and the subset of histograms from $\mathcal{H}^*$ form a quasi-$n$-cycle, a nontrivial solution is guaranteed if all of the $m$-cycles, $m < n$, which make it up are satisfied. To prove the only if statement, note that if any $n$-cycle is not satisfied, there exists a variable that is forced to equal zero. □

For the case $r = 1$, $K = 1$ (items may be classified as 'success' or 'failure'), Theorem 2.4.1 can be used to prove the coherency of any collection of predictive probabilities.

Let $\mathcal{H}'$ be the set of all possible conditioning histograms.

**Theorem 2.4.2** *In the case $r = 1$, $K = 1$, if you assert your predictive probabilities in (2.5) for all $H \in \mathcal{H}'$ to be strictly positive, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Theorem 2.4.1 will prove the statement of the theorem if it can be shown that $\mathcal{H}'$ is linked and that no $n$-cycles are formed by histograms from this set. When $r = K = 1$ the number of possible types is ${}^{r+K}C_K = {}^{2}C_1 = 2$ and the number of possible histograms is ${}^{(N+{}^{r+K}C_K-1)}C_N = {}^{N+1}C_N = N+1$. Consider $Hu$, the $u^{\text{th}}$ histogram in order from $\mathcal{H}'$, where $u \in \{1,\ldots,N+1\}$. Letting $x_1$ and $x_2$ denote the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type 1 and type 2 in $Hu$, respectively,

$$\begin{aligned} x_1 &= N+1-u \\ x_2 &= u-1\,. \end{aligned}$$

Then the variables in the equations associated with the histogram ordered immediately before $Hu$ $(u \neq 1)$ represent

$$P(N+3-u, u-2) \text{ and } P(N+2-u, u-1),$$

the variables in the equations associated with $Hu$ represent

$$P(N+2-u, u-1) \text{ and } P(N+1-u, u)$$

and the variables in the equations associated with the histogram ordered immediately after $Hu$ $(u \neq N+1)$ represent

$$P(N+1-u, u) \text{ and } P(N-u, u+1).$$

Clearly, $H1$ is linked to $H2$ which is also linked to $H3$ and so on, until $HN$ is linked to the last histogram. Hence $\mathcal{H}'$ is linked, with all of the histograms forming a 'chain' and no possibility of $n$-cycles being present. □

**Corollary 2.4.3** *The solution to the system of equations in Theorem 2.4.2 may be expressed*

$$q_a = {}^{N+1}C_{a-1} \prod_{u=1}^{a-1} \left[ \frac{1-p_{1,Hu}}{p_{1,Hu}} \right] q_1,$$

*or, equivalently,*

$$P(N+2-a, a-1)) = {}^{N+1}C_{a-1} \prod_{u=1}^{a-1} \left[ \frac{1-p_{1,Hu}}{p_{1,Hu}} \right] P(N+1, 0),$$

$\forall a \in \{2, \ldots, N+1\}$, *where* $Hu$, $u = 1, \ldots, N+1$, *is the* $u^{th}$ *histogram in order from* $\mathcal{H}'$.

PROOF: Variable $q_a$ is the second of the two variables in the equations associated with the histogram ordered immediately before $Ha$. By applying Theorem 2.3.2 to this histogram, $q_a$ can be expressed in terms of $q_{a-1}$. Now $q_{a-1}$ is the second of the two variables in the equations associated with the histogram ordered two before $Ha$. By applying Theorem 2.3.2 to this histogram, $q_{a-1}$ can be expressed in terms of $q_{a-2}$. Continuing in this fashion, $q_2$ is eventually expressed in terms of $q_1$ by

applying Theorem 2.3.2 to $H1$. The expression created is of the form

$$\begin{aligned} q_a &= \prod_{u=1}^{a-1} \left[ \frac{f_{1,Hu} p_{2,Hu}}{f_{2,Hu} p_{1,Hu}} \right] q_1 \\ &= \prod_{u=1}^{a-1} \left[ \frac{\left(\frac{N+2-u}{N+1}\right) p_{2,Hu}}{\left(\frac{u}{N+1}\right) p_{1,Hu}} \right] q_1 \\ &= \prod_{u=1}^{a-1} \left[ \left( \frac{N+2-u}{u} \right) \frac{p_{2,Hu}}{p_{1,Hu}} \right] q_1 \\ &= {}^{N+1}C_{a-1} \prod_{u=1}^{a-1} \left[ \frac{1-p_{1,Hu}}{p_{1,Hu}} \right] q_1 \end{aligned}$$

and Theorem 2.4.2 shows that this is the unique representation of $q_a$ in terms of $q_1$. □

The results given in Theorem 2.4.2 and Corollary 2.4.3 also appear in Lad, Deely and Piesse [57, 58] where they were derived in a different manner.

Having established Theorem 2.4.1, we can now use the theory of $n$-cycles to analyse the coherency of various methods and strategies for specifying the predictive probabilities in (2.5).

# Chapter 3

# Naive Probability Estimates

It seems natural to start by considering intuitive ways of specifying predictive probabilities and the situations for which you would be prepared to assert them.

## 3.1 The Frequency Mimicking Approach

One such naive strategy would be to assert

$$P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_N) = H)\right) = \frac{x_t}{N}, \qquad t = 1, \ldots, {}^{r+K}C_K\,, \tag{3.1}$$

where $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$. This amounts to a straight-out frequency mimicking approach.

**Definition 3.1.1** *A histogram, $H$, is said to be **strictly positive** if at least one of each of the ${}^{r+K}C_K$ types has been observed among the outcomes of the first $N$ groups. That is, if $min(x_1, \ldots, x_{r+KC_K}) > 0$ where $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$.*

Usually one would only be willing to assert probabilities of the form (3.1) when the conditioning histogram, $H$, is strictly positive. Obviously, this is only possible when $N \geq {}^{r+K}C_K$. Let $\mathcal{H}_{SP}$ be the set consisting of all strictly positive histograms. Note then that

$$p_{t,H} = \frac{x_t}{N} > 0, \qquad \forall\, t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, H \in \mathcal{H}_{SP}\,.$$

**Definition 3.1.2** *The* ***distance*** *between any two histograms, $h$ and $\mathfrak{h}$, is defined to be*

$$d(h,\mathfrak{h}) = \sum_{t=1}^{r+K C_K} |x_t - y_t| ,$$

*where $x_t$ and $y_t$, $t = 1,\ldots,{}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type $t$ in $h$ and $\mathfrak{h}$, respectively.*

Note that the distance between any two histograms, $h$ and $\mathfrak{h}$, is a nonnegative number and must be even for the following reason. Let $x_t$ and $y_t$, $t = 1,\ldots,{}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type $t$ in $h$ and $\mathfrak{h}$, respectively. Let $S$, $T$ and $U$ be the subsets of indices from $\left\{1,\ldots,{}^{r+K}C_K\right\}$ for which $x_t$ is greater than, less than and equal to $y_t$, respectively. Then

$$\begin{aligned} d(h,\mathfrak{h}) &= \sum_{t=1}^{r+K C_K} |x_t - y_t| \\ &= \sum_{\substack{t=1 \\ t\in S}}^{r+K C_K} (x_t - y_t) + \sum_{\substack{t=1 \\ t\in T}}^{r+K C_K} (y_t - x_t) + \sum_{\substack{t=1 \\ t\in U}}^{r+K C_K} 0 \\ &= 2 \sum_{\substack{t=1 \\ t\in S}}^{r+K C_K} (x_t - y_t) \end{aligned}$$

due to the fact that

$$\sum_{t=1}^{r+K C_K} x_t = \sum_{t=1}^{r+K C_K} y_t = N .$$

**Lemma 3.1.1** *The set $\mathcal{H}_{SP}$ of all strictly positive histograms is linked.*

PROOF: Take any two distinct histograms, $h$ and $\mathfrak{h} \in \mathcal{H}_{SP}$. Let $y_t$, $t = 1,\ldots,{}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type $t$ in $\mathfrak{h}$. Consider the following algorithm.

i. Let $Hu = h$.

ii. If $d(Hu,\mathfrak{h}) = 0$, stop. Otherwise, let $x_t$, $t = 1,\ldots,{}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type $t$ in $Hu$. Since

$$\sum_{t=1}^{r+K C_K} x_t = \sum_{t=1}^{r+K C_K} y_t = N ,$$

there exist $s_1, s_2 \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $x_{s_1} > y_{s_1}$, $x_{s_2} < y_{s_2}$. Link $Hu$ to $Hv$, where $Hv$ is defined to be the histogram that has one less type $s_1$ and one more type $s_2$ than $Hu$ does. These histograms may be written

$$\begin{array}{cc} x_1, \ldots, x_{s_1}, \ldots, x_{s_2}, \ldots, x_{r+K C_K} & Hu \\ x_1, \ldots, x_{s_1} - 1, \ldots, x_{s_2} + 1, \ldots, x_{r+K C_K} & Hv \\ y_1, \ldots, y_{s_1}, \ldots, y_{s_2}, \ldots, y_{r+K C_K} & \mathfrak{h}, \end{array}$$

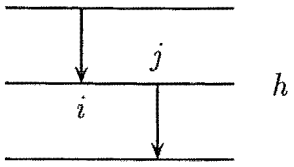
assuming, without loss of generality, that $s_1 < s_2$. Obviously, $d(Hv, \mathfrak{h}) = d(Hu, \mathfrak{h}) - 2$.

iii. Let $Hu = Hv$. Go to Step ii.

The algorithm is guaranteed to terminate because the distance of the current histogram from $\mathfrak{h}$ is being reduced each time through and is bounded below by 0. Since $h$ and $\mathfrak{h}$ are distinct, $d(h, \mathfrak{h}) \neq 0$ so that the algorithm will not stop the first time Step ii is encountered. If it stops the second time Step ii is encountered, then $h$ and $\mathfrak{h}$ are directly linked. Otherwise, $h$ and $\mathfrak{h}$ are linked indirectly through the sequence of histograms $\{Hv\}$. The only point that needs to be verified is that all of these 'intermediate' histograms are also members of $\mathcal{H}_{SP}$, *i.e.*, strictly positive histograms. This is evident for the following reason. For any intermediate histogram, the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, it has of any given type lies in the closed interval formed by the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, that $h$ has of that type, and the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, that $\mathfrak{h}$ has of that type. The fact that $h$ and $\mathfrak{h}$ are strictly positive histograms ensures that this interval excludes zero. □

**Theorem 3.1.2** *If you assert your predictive probabilities using (3.1) for all $H \in \mathcal{H}_{SP}$, the set of strictly positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Consider a given $n$-cycle formed by histograms from $\mathcal{H}_{SP}$. As we have already discovered, each occurrence




in the $n$-cycle contributes $p_{i,h}/p_{j,h}$ to the relevant ratio of probabilities. Let $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $h$. Now

$$\begin{aligned} \frac{p_{i,h}}{p_{j,h}} &= \frac{x_i/N}{x_j/N} \\ &= \frac{x_i}{x_j} \\ &= \frac{(N+1)f_{i,h} - 1}{(N+1)f_{j,h} - 1}. \end{aligned} \tag{3.2}$$

For exactly the same reason that the $f$ terms cancel to leave an overall ratio of probabilities only (see the discussion following Equation (2.15) of Chapter 2), Equation (3.2) shows that $p_{i,h}/p_{j,h}$ will cancel with corresponding contributions from other histograms in the $n$-cycle. Hence the overall ratio of predictive probabilities equals 1 and the $n$-cycle is satisfied. Lemma 3.1.1 and the statement immediately preceding Definition 3.1.2 then allow Theorem 2.4.1 to complete the proof. □

Having established Theorem 3.1.2, it would be useful to write down the exact expression for the solution to the system of equations referred to therein. The variables involved in these equations are exactly those components of the vector $\underline{q}$ that represent your prevision for outcomes of the $N + 1$ groups including among them some subset of $N$ that form a strictly positive histogram.

**Corollary 3.1.3** *The solution to the system of equations in Theorem 3.1.2 may be expressed*

$$q_{h(t)} = \frac{N + 2 - {}^{r+K}C_K}{(x_t + 1)\prod_{\substack{m=1 \\ m \neq t}}^{{}^{r+K}C_K} x_m} q_{\mathfrak{h}(1)},$$

*or, equivalently,*

$$P(x_1, \ldots, x_t + 1, \ldots, x_{{}^{r+K}C_K}) = \frac{N + 2 - {}^{r+K}C_K}{(x_t + 1)\prod_{\substack{m=1 \\ m \neq t}}^{{}^{r+K}C_K} x_m} P\left(N + 2 - {}^{r+K}C_K, 1, \ldots, 1\right),$$

$\forall\, t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, h \in \mathcal{H}_{SP}$, *where* $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, *denotes the number of* $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, *of type* $m$ *in* $h$, *and* $\mathfrak{h} \in \mathcal{H}_{SP}$ *is the histogram consisting of* $\left(N + 1 - {}^{r+K}C_K\right)$ $\underline{Y}^{(i)}$ *of type 1, and one* $\underline{Y}^{(i)}$ *of each of type 2, ..., type* ${}^{r+K}C_K$.

PROOF: Consider the following algorithm.

i. Use Theorem 2.3.2 to express $q_{h(t)}$ in terms of $q_{h(1)}$,

$$q_{h(t)} = \frac{f_{1,h}p_{t,h}}{f_{t,h}p_{1,h}} q_{h(1)} .$$

ii. Let $Hu = h$.

iii. If $d(Hu, \mathfrak{h}) = 0$, stop. Otherwise, let $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $m$ in $Hu$. The fact that $Hu \in \mathcal{H}_{SP}$ then implies that $x_1 < N + 1 - {}^{r+K}C_K$ and there exists $s \in \left\{2, \ldots, {}^{r+K}C_K\right\}$ defined to be the smallest member of the set of indices of those components of $Hu$ that exceed 1. Hence $x_s > 1$. Link $Hu$ to $Hv$, where $Hv$ is defined to be the histogram that has one less type $s$ and one more type 1 than $Hu$ does. These histograms may be written

$$\begin{array}{cc} N + 1 - {}^{r+K}C_K, 1, \ldots, 1 & \mathfrak{h} \\ x_1 + 1, 1, \ldots, 1, x_s - 1, \ldots, x_{r+K C_K} & Hv \\ x_1, 1, \ldots, 1, x_s, \ldots, x_{r+K C_K} & Hu . \end{array}$$

Obviously, $q_{Hu(1)} = q_{Hv(s)}$ (recall the definition of $q_{H(t)}$ in Equation (2.10) of Chapter 2) and $d(Hv, \mathfrak{h}) = d(Hu, \mathfrak{h}) - 2$.

iv. Use Theorem 2.3.2 to express $q_{Hv(s)}$ in terms of $q_{Hv(1)}$,

$$q_{Hv(s)} = \frac{f_{1,Hv}p_{s,Hv}}{f_{s,Hv}p_{1,Hv}} q_{Hv(1)} .$$

v. Let $Hu = Hv$. Go to Step iii.

The algorithm is guaranteed to terminate because the distance of the current histogram from $\mathfrak{h}$ is being reduced each time through and is bounded below by 0. If $h$ is the same histogram as $\mathfrak{h}$, $d(h, \mathfrak{h}) = 0$ so that the algorithm will stop the first time Step iii is encountered. In this case, by Theorem 2.3.2,

$$\begin{aligned} q_{h(t)} &= q_{\mathfrak{h}(t)} \\ &= \frac{f_{1,\mathfrak{h}}p_{t,\mathfrak{h}}}{f_{t,\mathfrak{h}}p_{1,\mathfrak{h}}} q_{\mathfrak{h}(1)} \end{aligned}$$

$$= \begin{cases} q_{\mathfrak{h}(1)}, & t = 1 \\ \dfrac{\left(\frac{N+2-{}^{r+K}C_K}{N+1}\right)\left(\frac{1}{N}\right)}{\left(\frac{2}{N+1}\right)\left(\frac{N+1-{}^{r+K}C_K}{N}\right)} q_{\mathfrak{h}(1)}, & t \neq 1 \end{cases}$$

$$= \frac{N+2-{}^{r+K}C_K}{(x_t+1)\prod_{\substack{m=1 \\ m\neq t}}^{{}^{r+K}C_K} x_m} q_{\mathfrak{h}(1)}\,, \tag{3.3}$$

where $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $m$ in $\mathfrak{h}$. Otherwise, let $n$ be the number of times Step iii is encountered, *not* including the last time (at which point the algorithm stops) and let the sequence of histograms linking $h$ to $\mathfrak{h}$, $\{Hv\} = \{Hv_1, \ldots, Hv_n\}$. Note that $Hv_n = \mathfrak{h}$. Reverting to the notation where $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $m$ in $h$, these histograms may be written

$$\begin{array}{cc} N+1-{}^{r+K}C_K, 1, \ldots, 1 & Hv_n = \mathfrak{h} \\ N-{}^{r+K}C_K, 1, \ldots, 1, x_{s_{n-1}} = 2, 1, \ldots, 1 & Hv_{n-1} \\ \vdots & \\ x_1+1, 1, \ldots, 1, x_{s_0}-1, \ldots, x_{r+K C_K} & Hv_1 \\ x_1, 1, \ldots, 1, x_{s_0}, \ldots, x_{r+K C_K} & h\,, \end{array}$$

where $s_{i-1}$, $i = 1, \ldots, n$, is the value of $s$ in Step iii of the algorithm at the $i^{\text{th}}$ time Step iii is encountered. That is, $s_0$ is the smallest member of the set of indices of those components of $h$ that exceed 1, and $s_{i-1}$, $i = 2, \ldots, n$, is the smallest member of the set of indices of those components of $Hv_{i-1}$ that exceed 1. Thus, the expression derived by the implementation of the algorithm is

$$q_{h(t)} = \frac{f_{1,h}p_{t,h}}{f_{t,h}p_{1,h}} \prod_{i=1}^{n} \left[\frac{f_{1,Hv_i}p_{s_{i-1},Hv_i}}{f_{s_{i-1},Hv_i}p_{1,Hv_i}}\right] q_{\mathfrak{h}(1)}\,.$$

Let

$$p'_{m,H} = Np_{m,H}, \quad f'_{m,H} = (N+1)f_{m,H}, \qquad \forall\, m \in \left\{1, \ldots, {}^{r+K}C_K\right\}, H \in \mathcal{H}_{SP}\,.$$

Then

$$f'_{1,h} = p'_{1,Hv_1}$$

and

$$f'_{1,Hv_i} = p'_{1,Hv_{i+1}}, \qquad i = 1, \ldots, n-1\,.$$

Hence,

$$\begin{aligned}\frac{f'_{1,h}}{p'_{1,h}}\prod_{i=1}^{n}\frac{f'_{1,Hv_i}}{p'_{1,Hv_i}} &= \frac{f'_{1,Hv_n}}{p'_{1,h}}\\ &= \frac{f'_{1,h}}{p'_{1,h}}\\ &= \frac{N+2-{}^{r+K}C_K}{x_1}.\end{aligned}$$

Consider any $s_I \in \{s_0,\ldots,s_{n-1}\}$. Let $i_{min}$ and $i_{max}$ be the smallest and largest indices from $\{0,\ldots,n-1\}$, respectively, for which $s_i = s_I$. If $i_{min} = i_{max}$, it must be that $x_{s_I} = 2$. Then

$$\begin{aligned}\prod_{i=i_{min}}^{i_{max}}\frac{p'_{s_{i-1},Hv_i}}{f'_{s_{i-1},Hv_i}} &= \frac{p'_{s_{i_{min}-1},Hv_{i_{min}}}}{f'_{s_{i_{min}-1},Hv_{i_{min}}}}\\ &= \frac{1}{2}\\ &= \frac{1}{x_{s_I}}.\end{aligned}$$

Otherwise,

$$p'_{s_{i-1},Hv_i} = f'_{s_{i-1},Hv_{i+1}}, \qquad i = i_{min}+1,\ldots,i_{max},$$

and

$$\begin{aligned}\prod_{i=i_{min}}^{i_{max}}\frac{p'_{s_{i-1},Hv_i}}{f'_{s_{i-1},Hv_i}} &= \frac{p'_{s_{i_{min}-1},Hv_{i_{min}}}p'_{s_{i_{max}-1},Hv_{i_{max}}}}{f'_{s_{i_{min}-1},Hv_{i_{min}}}f'_{s_{i_{min}},Hv_{i_{min}+1}}}\\ &= \frac{p'_{s_{i_{min}-1},Hv_{i_{min}}}p'_{s_I,Hv_{i_{max}}}}{f'_{s_{i_{min}-1},Hv_{i_{min}}}f'_{s_I,Hv_{i_{min}+1}}}\\ &= \frac{1\,.\,2}{2\,.\,x_{s_I}}\\ &= \frac{1}{x_{s_I}}.\end{aligned}$$

In either case it follows that

$$\prod_{i=1}^{n}\frac{p'_{s_{i-1},Hv_i}}{f'_{s_{i-1},Hv_i}} = \prod_{\substack{m=2\\ m\in S}}^{r+K C_K}\frac{1}{x_m},$$

where $S$ is now defined to be the set of indices of those components of $h$ that exceed 1. Hence, $m \in S \Leftrightarrow x_m > 1$. Now

$$\begin{aligned}
\frac{f_{1,h}p_{t,h}}{f_{t,h}p_{1,h}}\prod_{i=1}^{n}\left[\frac{f_{1,Hv_i}p_{s_{i-1},Hv_i}}{f_{s_{i-1},Hv_i}p_{1,Hv_i}}\right] &= \frac{f'_{1,h}p'_{t,h}}{f'_{t,h}p'_{1,h}}\prod_{i=1}^{n}\left[\frac{f'_{1,Hv_i}p'_{s_{i-1},Hv_i}}{f'_{s_{i-1},Hv_i}p'_{1,Hv_i}}\right] \\
&= \frac{f'_{1,h}}{p'_{1,h}}\prod_{i=1}^{n}\left[\frac{f'_{1,Hv_i}}{p'_{1,Hv_i}}\right]\frac{p'_{t,h}}{f'_{t,h}}\prod_{i=1}^{n}\left[\frac{p'_{s_{i-1},Hv_i}}{f'_{s_{i-1},Hv_i}}\right] \\
&= \frac{\left(N+2-{}^{r+K}C_K\right)}{x_1}\frac{x_t}{(x_t+1)}\prod_{\substack{m=2\\ m\in S}}^{{}^{r+K}C_K}\frac{1}{x_m} \\
&= \frac{x_t\left(N+2-{}^{r+K}C_K\right)}{(x_t+1)x_1\prod_{\substack{m=2\\ m\in S}}^{{}^{r+K}C_K}x_m} \\
&= \frac{x_t\left(N+2-{}^{r+K}C_K\right)}{(x_t+1)\prod_{m=1}^{{}^{r+K}C_K}x_m} \\
&= \frac{N+2-{}^{r+K}C_K}{(x_t+1)\prod_{\substack{m=1\\ m\neq t}}^{{}^{r+K}C_K}x_m}. \qquad (3.4)
\end{aligned}$$

Hence, by (3.3) and (3.4),

$$q_{h(t)} = \frac{N+2-{}^{r+K}C_K}{(x_t+1)\prod_{\substack{m=1\\ m\neq t}}^{{}^{r+K}C_K}x_m}q_{\mathfrak{h}(1)},$$

$\forall\, t \in \left\{1,\ldots,{}^{r+K}C_K\right\}, h \in \mathcal{H}_{SP}$, where $x_m$, $m = 1,\ldots,{}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1,\ldots,N$, of type $m$ in $h$. Theorem 3.1.2 shows that this is the unique representation of $q_{h(t)}$ in terms of $q_{\mathfrak{h}(1)}$. □

Theorem 3.1.2 and Corollary 3.1.3 also occur in Lad, Deely and Piesse [57] for the case $r = 1$, and in Lad,Deely and Piesse [58] for the case $r = K = 1$. However, in these references the proof of a coherent *nontrivial* solution relies on the ability to derive the exact expression for the solution to the coherency induced equations. This approach, which may not always be computationally practical, is not the only means of deciding whether or not a nontrivial solution exists, now that the notion of $n$-cycles within the equations has been introduced and their structure exploited.

The practical implication of Theorem 3.1.2 is that a strategy for estimating predictive probabilities that would employ a frequency mimicking approach, given any strictly positive histogram, is indeed coherent. More importantly, it is coherent without the need for the concomitant assertion of zero probability of the first $N$ groups producing a strictly positive histogram. This result may seem appealing,

especially to an objectivist frequentist statistician (who would like exclusively to use observed frequencies as predictive probability assertions), as it is comforting to know that such a naive, intuitive strategy is coherent. However, it is shown in [57] (for $r = 1$) and in [58] (for $r = K = 1$) that there are at least three unattractive concomitant assertions required of you if you would employ this frequency mimicking approach. They are recapitulated here.

i. Although such a strategy may seem feasible when the number of groups, $N$, making up the conditioning histogram is large, coherency requires that observed frequencies must also represent your predictive probabilities when the number of groups is *small*. Specifically, for any $M < N$,

$$P\left(\underline{Y}^{(M+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_M) = H)\right) = \frac{x_t}{M}, \qquad t = 1, \ldots, {}^{r+K}C_K\,,$$

where $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, M$, of type $t$ in $H \in \mathcal{H}_{SP}$.

ii. Any two strictly positive histograms formed by *all* $N+1$ groups that are permutations of one another must be accorded identical probabilities (an obvious consequence of Corollary 3.1.3).

iii. The *systematic* strategy to use observed frequencies as predictive probabilities (given any strictly positive histogram) for *any* size of $N$ would only be coherent along with the assertion

$$\sum_{H \in \mathcal{H}_{SP}} P(H(\underline{Y}_N) = H) = 0$$

for every value of $N$. Thus, an assertion of positive probability that all types of $\underline{Y}^{(i)}$ will be observed within the outcomes of a finite number of groups would require some adjustment to your specified predictive probabilities.

None of these assertions that are required to accompany the frequency mimicking approach to inference seems warranted in many real problems. Thus, the practical statistician would be driven away from systematic use of this approach, except perhaps in very small scale problems.

## 3.2 Generalising the Frequency Mimicking Approach

The suggestion of an adjustment to the straight-out frequency mimicking approach could be carried out in many different ways. One obvious possibility would be to assert

$$P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_N) = H)\right) = \frac{x_t + c}{N + ({}^{r+K}C_K)c}, \quad t = 1, \ldots, {}^{r+K}C_K\,, \quad (3.5)$$

where $c > 0, c \in \mathbb{R}$ and $x_t, t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$. Similarly, another possibility would be to assert

$$P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_N) = H)\right) = \frac{x_t + c_t}{N + \sum_{s=1}^{r+K C_K} c_s}, \quad t = 1, \ldots, {}^{r+K}C_K\,, \quad (3.6)$$

where $c_t > 0, c_t \in \mathbb{R}, t = 1, \ldots, {}^{r+K}C_K$, and $x_t, t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$.

The astute reader may have noticed a similarity between the expressions in (3.5) and (3.6) and a posterior expectation associated with a Dirichlet distribution (symmetric in the case of (3.5)). This relationship will be explored in Chapter 5.

**Theorem 3.2.1** *If you assert your predictive probabilities using (3.5) for all $H \in \mathcal{H}_{SP}$, the set of strictly positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Follow the proof of Theorem 3.1.2 but replace Equation (3.2) with

$$\begin{aligned} \frac{p_{i,h}}{p_{j,h}} &= \frac{(x_i + c)\Big/\big(N + ({}^{r+K}C_K)c\big)}{(x_j + c)\Big/\big(N + ({}^{r+K}C_K)c\big)} \\ &= \frac{x_i + c}{x_j + c} \\ &= \frac{(N+1)f_{i,h} + c - 1}{(N+1)f_{j,h} + c - 1}. \end{aligned}$$

□

**Corollary 3.2.2** *The solution to the system of equations in Theorem 3.2.1 may be expressed*

$$q_{h(t)} = P(x_1, \ldots, x_t + 1, \ldots, x_{r+{}^{K}C_K}) =$$

$$\frac{\Gamma(N+3-{}^{r+K}C_K)}{\Gamma(c+1)^{({}^{r+K}C_K-1)}\Gamma(N+2-{}^{r+K}C_K+c)} \frac{\Gamma(x_t+1+c)}{\Gamma(x_t+2)} \prod_{\substack{m=1 \\ m\neq t}}^{{}^{r+K}C_K} \left[\frac{\Gamma(x_m+c)}{\Gamma(x_m+1)}\right] q_{\mathfrak{h}(1)},$$

$\forall\, t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, h \in \mathcal{H}_{SP}$, *where* $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, *denotes the number of* $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, *of type* $m$ *in* $h$, *and* $\mathfrak{h} \in \mathcal{H}_{SP}$ *is the histogram consisting of* $\left(N+1-{}^{r+K}C_K\right)$ $\underline{Y}^{(i)}$ *of type 1, and one* $\underline{Y}^{(i)}$ *of each of type* $2, \ldots,$ *type* ${}^{r+K}C_K$ *so that* $q_{\mathfrak{h}(1)} = P\left(N+2-{}^{r+K}C_K, 1, \ldots, 1\right)$.

PROOF: Similar to the proof of Corollary 3.1.3. □

**Theorem 3.2.3** *If you assert your predictive probabilities using (3.6) for all* $H \in \mathcal{H}_{SP}$, *the set of strictly positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Follow the proof of Theorem 3.1.2 but replace Equation (3.2) with

$$\begin{aligned} \frac{p_{i,h}}{p_{j,h}} &= \frac{(x_i + c_i) \Big/ \left(N + \sum_{s=1}^{r+K C_K} c_s\right)}{(x_j + c_j) \Big/ \left(N + \sum_{s=1}^{r+K C_K} c_s\right)} \\ &= \frac{x_i + c_i}{x_j + c_j} \\ &= \frac{(N+1) f_{i,h} + c_i - 1}{(N+1) f_{j,h} + c_j - 1}. \end{aligned}$$

□

**Corollary 3.2.4** *The solution to the system of equations in Theorem 3.2.3 may be expressed*

$$q_{h(t)} = P(x_1, \ldots, x_t + 1, \ldots, x_{r+{}^{K}C_K}) =$$

$$\frac{\Gamma(N+3-{}^{r+K}C_K)}{\prod_{s=2}^{{}^{r+K}C_K}[\Gamma(c_s+1)]\Gamma(N+2-{}^{r+K}C_K+c_1)} \frac{\Gamma(x_t+1+c_t)}{\Gamma(x_t+2)} \prod_{\substack{m=1 \\ m\neq t}}^{{}^{r+K}C_K} \left[\frac{\Gamma(x_m+c_m)}{\Gamma(x_m+1)}\right] q_{\mathfrak{h}(1)},$$

$\forall\, t \in \left\{1, \ldots, {}^{r+K}C_K\right\}, h \in \mathcal{H}_{SP}$, *where* $x_m$, $m = 1, \ldots, {}^{r+K}C_K$, *denotes the number of* $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, *of type* $m$ *in* $h$, *and* $\mathfrak{h} \in \mathcal{H}_{SP}$ *is the histogram consisting of*

$\left(N+1-{}^{r+K}C_K\right)\underline{Y}^{(i)}$ *of type 1, and one* $\underline{Y}^{(i)}$ *of each of type* $2,\ldots,$ *type* ${}^{r+K}C_K$ *so that* $q_{\mathfrak{h}(1)} = P\left(N+2-{}^{r+K}C_K, 1, \ldots, 1\right)$.

PROOF: Similar to the proof of Corollary 3.1.3. □

Note that predictive probabilities of the form (3.5) or (3.6) will always be strictly positive, so that Theorems 3.2.1 and 3.2.3 hold equally well if they refer to a strategy that asserts these probabilities for *all* possible conditioning histograms. (The fact that the set of all possible histograms, $\mathcal{H}'$, is linked follows from application of the algorithm described in the proof of Lemma 3.1.1 to this set.)

Also, as the constant, $c$, in (3.5) increases, the predictive probabilities are shrunk away from their raw observed frequencies towards

$$p_{t,H} = \frac{1}{{}^{r+K}C_K}, \qquad t = 1, \ldots, {}^{r+K}C_K\,,$$

in the extreme ($c \to \infty$).

Again, the practical implications of Theorems 3.2.1 and 3.2.3 are that the strategies given by (3.5) and (3.6) represent coherent ways of specifying your predictive probabilities, either when employed only given a strictly positive histogram or when given any conditioning histogram. Also, they are coherent without the need for the concomitant assertion of zero probability of the first $N$ groups producing such histograms. It remains to ascertain whether or not these strategies require annoying concomitant assertions as does the unadjusted frequency mimicking approach. Referring to these as previously itemised on page 49, it can be shown that an equivalent of Assertion i is still required. That is, if you assert your predictive probabilities using (3.5) (which contains (3.6) as a special case), for all $H \in \mathcal{H}_{SP}$, then coherency requires that (3.5) must also represent your predictive probabilities when the number of groups is smaller than $N$. Assertion ii is required when employing the strategy given by (3.5), but not necessarily when employing the strategy given by (3.6). (This follows from Corollaries 3.2.2 and 3.2.4.) Possibly of most interest, and relevance, is that the equivalent form of Assertion iii is no longer required for the strategy given by (3.5). This will be proven in Theorem 3.2.17 as the culmination of numerous necessary side-results that will be established along the way. Presumably the strategy given by (3.6) also has this advantage.

The proof of Assertion iii for the straight-out frequency mimicking approach was accomplished in [57] along the following lines.

- Fix $M \leq N$, and consider $P(H(\underline{Y}_M) = h(M))$ where $h(M)$ is a strictly positive histogram.
- Express $P(H(\underline{Y}_M) = h(M))$ as a function of $P(H(\underline{Y}_{N+1}) = \mathfrak{h})$, where $\mathfrak{h}$ is the histogram consisting of $\left(N + 2 - {}^{r+K}C_K\right) \underline{Y}^{(i)}$ of type 1, and one $\underline{Y}^{(i)}$ of each of type 2, ..., type ${}^{r+K}C_K$.
- Show that this expression for $P(H(\underline{Y}_M) = h(M))$ converges to 0 as $N$ increases.
- Note that because there are only a finite number of positive histograms, $h(M)$, your probability for achieving a strictly positive histogram from the first $M$ groups must therefore be zero if you would employ the frequency mimicking strategy for *all* $N$.

Let us now return our attention to the case where your predictive probabilities are asserted using (3.5), given a strictly positive histogram, and proceed to follow the steps outlined above.

Fix $M \leq N$, and consider $P(H(\underline{Y}_M) = h(M))$ where $h(M) \in \mathcal{H}_{SP}$. Let $h_t(N+1)$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N+1$, of type $t$ in histogram $H(\underline{Y}_{N+1})$. For readability, let $R = {}^{r+K}C_K$. Then, by Corollary 3.2.2,

$$\begin{aligned} & P(H(\underline{Y}_M) = h(M)) \\ = & \sum{}^{*} \left[ \frac{{}^{h(N+1)}C_{h(M)}}{{}^{N+1}C_M} \frac{\Gamma(N+3-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2-R+c)} \prod_{t=1}^{R} \left[ \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] p^* \right], \end{aligned}$$

where the summation $\sum^*$ runs over all $H(\underline{Y}_{N+1}) > h(M) \in \mathcal{H}_{SP}$ and $p^*$ denotes $P\left(N + 2 - {}^{r+K}C_K, 1, \ldots, 1\right)$. Hence,

$$\begin{aligned} & P(H(\underline{Y}_M) = h(M)) \\ = & \sum{}^{*} \left[ \frac{{}^{h(N+1)}C_{h(M)}}{{}^{N+1}C_M} \frac{\Gamma(N+3-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2-R+c)} \prod_{t=1}^{R} \left[ \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] p^* \right] \\ = & \frac{\Gamma(N+3-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2-R+c)} p^* \sum{}^{*} \left[ \frac{{}^{h(N+1)}C_{h(M)}}{{}^{N+1}C_M} \prod_{t=1}^{R} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \\ = & \frac{\Gamma(N+3-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2-R+c)} {}^{M}C_{h(M)}\, p^* \times \\ & \sum{}^{*} \left[ \frac{{}^{N+1-M}C_{h(N+1)-h(M)}}{{}^{N+1}C_{h(N+1)}} \prod_{t=1}^{R} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(N+3-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2-R+c)}{}^{M}C_{h(M)}\,p^* \times \\
&\qquad \sum{}^{*}\left[{}^{N+1-M}C_{h(N+1)-h(M)}\frac{\prod_{t=1}^{R}\Gamma(h_t(N+1)+1)}{\Gamma(N+2)}\prod_{t=1}^{R}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right] \\
&= \frac{\Gamma(N+3-R)\Gamma(N+2+Rc-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2)\Gamma(N+2-R+c)}{}^{M}C_{h(M)}\,p^* \times \\
&\qquad \sum{}^{*}\left[{}^{N+1-M}C_{h(N+1)-h(M)}\frac{\prod_{t=1}^{R}\Gamma(h_t(N+1)+c)}{\Gamma(N+2+Rc-R)}\right].
\end{aligned}
\tag{3.7}
$$

In order to analyse the behaviour of (3.7) for large $N$, it is necessary first to study the behaviour of $\sum^*$ and $p^*$ when $N$ is large. The following lemma is used in Lemma 3.2.6 which deals with $\sum^*$.

**Lemma 3.2.5** *If* $n \in \mathbb{N}$, $T \in \{1,\ldots,{}^{r+K}C_K - 1\}$, $c > 0, c \in \mathbb{R}$ *and* $y_t, d_t \in \mathbb{Z}$, $t = 1,\ldots,{}^{r+K}C_K$ *are such that*

$$
\begin{aligned}
y_t &> 0,\ t = 1,\ldots,{}^{r+K}C_K, & d_t &\geq 0,\ t = 1,\ldots,{}^{r+K}C_K, \\
\textstyle\sum_{t=1}^{r+K C_K} y_t &= M, & \textstyle\sum_{t=1}^{r+K C_K} d_t &= N+1-M,
\end{aligned}
$$

*then*

$$
\begin{aligned}
&\sum_{d_T=0}^{N+1-M-\sum_{t=1}^{T-1} d_t}\left[\frac{\Gamma(y_T+d_T+c)\Gamma\left(N+1-\sum_{t=1}^{T}(y_t+d_t)+nc\right)}{d_T!\left(N+1-M-\sum_{t=1}^{T} d_t\right)!}\right] \\
&= \frac{\Gamma(y_T+c)\Gamma\left(M-\sum_{t=1}^{T} y_t+nc\right)\Gamma\left(N+1-\sum_{t=1}^{T-1}(y_t+d_t)+(n+1)c\right)}{\left(N+1-M-\sum_{t=1}^{T-1} d_t\right)!\Gamma\left(M-\sum_{t=1}^{T-1} y_t+(n+1)c\right)}.
\end{aligned}
$$

PROOF:

$$
\begin{aligned}
&\text{LHS} \\
&= \Gamma(y_T+c)\Gamma\Big(M-\textstyle\sum_{t=1}^{T} y_t+nc\Big) \times \\
&\qquad \sum_{d_T=0}^{N+1-M-\sum_{t=1}^{T-1} d_t}\left[\frac{\Gamma(y_t+d_t+c)}{d_t!\Gamma(y_t+c)}\frac{\Gamma\left(N+1-\sum_{t=1}^{T}(y_t+d_t)+nc\right)}{\left(N+1-M-\sum_{t=1}^{T} d_t\right)!\Gamma\left(M-\sum_{t=1}^{T} y_t+nc\right)}\right] \\
&= \Gamma(y_T+c)\Gamma\Big(M-\textstyle\sum_{t=1}^{T} y_t+nc\Big) \sum_{d_T=0}^{N+1-M-\sum_{t=1}^{T-1} d_t}\Big[\Big(\text{coeff. of } z^{d_T} \text{ in} \\
&\qquad (1-z)^{y_T+c}\Big)\Big(\text{coeff. of } z^{\left(N+1-M-\sum_{t=1}^{T} d_t\right)} \text{ in } (1-z)^{\left(M-\sum_{t=1}^{T} y_t+nc\right)}\Big)\Big] \\
&= \Gamma(y_T+c)\Gamma\Big(M-\textstyle\sum_{t=1}^{T} y_t+nc\Big) \times
\end{aligned}
$$

$$\begin{aligned}
&\left(\text{coeff. of } z^{\left(N+1-M-\sum_{t=1}^{T-1} d_t\right)} \text{ in } (1-z)^{\left(M-\sum_{t=1}^{T-1} y_t+(n+1)c\right)}\right) \\
= \; & \frac{\Gamma(y_T+c)\Gamma\left(M-\sum_{t=1}^{T} y_t+nc\right)\Gamma\left(N+1-\sum_{t=1}^{T-1}(y_t+d_t)+(n+1)c\right)}{\left(N+1-M-\sum_{t=1}^{T-1} d_t\right)!\,\Gamma\left(M-\sum_{t=1}^{T-1} y_t+(n+1)c\right)} \\
= \; & \text{RHS}
\end{aligned}$$

□

**Lemma 3.2.6** *For large $N$,*

$$\sum{}^{*}\left[{}^{N+1-M}C_{h(N+1)-h(M)}\frac{\prod_{t=1}^{r+K C_K}\Gamma(h_t(N+1)+c)}{\Gamma\left(N+2+({}^{r+K}C_K)c-{}^{r+K}C_K\right)}\right] \sim N^{{}^{r+K}C_K-1},$$

*where $c>0, c\in\mathbb{R}$ and the summation $\sum^{*}$ runs over all $H(\underline{Y}_{N+1})>h(M)\in\mathcal{H}_{SP}$.*

PROOF: Let $y_t$, $t=1,\ldots,{}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i=1,\ldots,N$, of type $t$ in $H(M)$, and let $d_t$, $t=1,\ldots,{}^{r+K}C_K$, denote the *difference* in the number of $\underline{Y}^{(i)}$, $i=1,\ldots,N$, of type $t$ in $H(\underline{Y}_{N+1})$ and $H(M)$. Let $R={}^{r+K}C_K$. Then recursive use of Lemma 3.2.5 shows that

$$\begin{aligned}
&\sum{}^{*}\left[{}^{N+1-M}C_{h_t(N+1)-h_t(M)}\frac{\prod_{t=1}^{R}\Gamma(h_t(N+1)+c)}{\Gamma(N+2+Rc-R)}\right] \\
= \; & \sum_{d_1=0}^{N+1-M}\sum_{d_2=0}^{N+1-M-d_1}\cdots\sum_{d_{R-1}=0}^{N+1-M-\sum_{t=1}^{R-2} d_t}\left[\frac{(N+1-M)!\prod_{t=1}^{R}\Gamma(y_t+d_t+c)}{\prod_{t=1}^{R}[d_t!]\Gamma(N+2+Rc-R)}\right] \\
= \; & \frac{(N+1-M)!}{\Gamma(N+2+Rc-R)}\sum_{d_1=0}^{N+1-M}\frac{\Gamma(y_1+d_1+c)}{d_1!}\sum_{d_2=0}^{N+1-M-d_1}\frac{\Gamma(y_2+d_2+c)}{d_2!}\times \\
& \cdots\times\sum_{d_{R-1}=0}^{N+1-M-\sum_{t=1}^{R-2} d_t}\left[\frac{\Gamma(y_{R-1}+d_{R-1}+c)\Gamma\left(N+1-\sum_{t=1}^{R-1}(y_t+d_t)+c\right)}{d_{R-1}!\left(N+1-M-\sum_{t=1}^{R-1} d_t\right)!}\right] \\
= \; & \frac{(N+1-M)!\Gamma(y_{R-1}+c)\Gamma\left(M-\sum_{t=1}^{R-1} y_t+c\right)}{\Gamma(N+2+Rc-R)\Gamma\left(M-\sum_{t=1}^{R-2} y_t+2c\right)}\sum_{d_1=0}^{N+1-M}\frac{\Gamma(y_1+d_1+c)}{d_1!}\times \\
& \cdots\times\sum_{d_{R-2}=0}^{N+1-M-\sum_{t=1}^{R-3} d_t}\left[\frac{\Gamma(y_{R-2}+d_{R-2}+c)\Gamma\left(N+1-\sum_{t=1}^{R-2}(y_t+d_t)+2c\right)}{d_{R-2}!\left(N+1-M-\sum_{t=1}^{R-2} d_t\right)!}\right] \\
& \vdots \\
= \; & \frac{(N+1-M)!\prod_{s=2}^{R-1}\left[\Gamma(y_s+c)\Gamma\left(M-\sum_{t=1}^{s} y_t+(R-s)c\right)\right]}{\Gamma(N+2+Rc-R)\prod_{s=1}^{R-2}\Gamma\left(M-\sum_{t=1}^{s} y_t+(R-s)c\right)}\times
\end{aligned}$$

$$\begin{aligned}
&\sum_{d_1=0}^{N+1-M}\left[\frac{\Gamma(y_1+d_1+c)\Gamma(N+1-y_1-d_1+(R-1)c)}{d_1!(N+1-M-d_1)!}\right] \\
= &\frac{(N+1-M)!\prod_{s=1}^{R-1}\left[\Gamma(y_s+c)\Gamma\left(M-\sum_{t=1}^{s}y_t+(R-s)c\right)\right]}{\Gamma(N+2+Rc-R)\prod_{s=1}^{R-2}\Gamma\left(M-\sum_{t=1}^{s}y_t+(R-s)c\right)}\times \\
&\frac{\Gamma(N+1+Rc)}{(N+1-M)!\Gamma(M+Rc)} \\
= &\frac{\Gamma(N+1+Rc)\prod_{s=1}^{R-1}[\Gamma(y_s+c)]\Gamma\left(M-\sum_{t=1}^{R-1}y_t+c\right)}{\Gamma(N+2+Rc-R)\Gamma(M+Rc)} \\
\sim &\ N^{R-1}, \quad \text{for large } N,
\end{aligned}$$

by Sterling's approximation (see [4]). □

Note that

$$\sum\nolimits^{+}P(H(\underline{Y}_{N+1}))+\sum\nolimits^{-}P(H(\underline{Y}_{N+1}))=1,$$

where the summation $\sum^{+}$ runs over all strictly positive histograms and the summation $\sum^{-}$ runs over all possible remaining histograms. Hence, by Corollary 3.2.2,

$$\frac{\Gamma\left(N+3-{}^{r+K}C_K\right)}{\Gamma(c+1)^{({}^{r+K}C_K-1)}\Gamma\left(N+2-{}^{r+K}C_K+c\right)}p^{*}\sum\nolimits^{+}\left[\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right] \tag{3.8}$$

$$+\sum\nolimits^{-}P(H(\underline{Y}_{N+1}))=1,$$

where $p^{*}=P\left(N+2-{}^{r+K}C_K,1,\ldots,1\right)$. Therefore, in order to analyse the behaviour of $p^{*}$ for large $N$, it is necessary first to study the behaviour of $\sum^{+}$ when $N$ is large. This is accomplished by finding bounds on the size of $\sum^{+}$, using Lemmas 3.2.7–3.2.14, and is summarised in Lemma 3.2.15.

**Lemma 3.2.7** *The expression*

$$\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(x_t+c)}{\Gamma(x_t+1)},$$

*subject to the constraints*

$$\begin{aligned}
x_t &\geq 0, \qquad t=1,\ldots,{}^{r+K}C_K, \\
\sum_{t=1}^{{}^{r+K}C_K}x_t &= N+1,
\end{aligned}$$

*where* $c\in\mathbb{R}$, *is minimised for* $0<c<1$, *and maximised for* $c>1$, *by*

$$x_1=x_2=\cdots=x_{{}^{r+K}C_K}=\frac{N+1}{{}^{r+K}C_K}.$$

PROOF: It suffices to find the extreme values of

$$\begin{aligned} f(\underline{x}) &\equiv \ln \prod_{t=1}^{r+{}^{K}C_K} \frac{\Gamma(x_t + c)}{\Gamma(x_t + 1)} \\ &= \sum_{t=1}^{r+{}^{K}C_K} \left[\ln \Gamma(x_t + c) - \ln \Gamma(x_t + 1)\right] . \end{aligned}$$

First, assume $0 < c < 1$. Now

$$\begin{aligned} \underline{g}(\underline{x}) &\equiv \nabla f(\underline{x}) \\ &= \begin{pmatrix} \psi(x_1 + c) - \psi(x_1 + 1) \\ \vdots \\ \psi(x_{r+{}^{K}C_K} + c) - \psi(x_{r+{}^{K}C_K} + 1) \end{pmatrix} , \end{aligned}$$

where $\psi(z) \equiv d(\ln \Gamma(z))/dz = \Gamma'(z)/\Gamma(z)$ is the so-called Digamma function. Let the equality constraint be

$$c(\underline{x}) = \sum_{t=1}^{r+{}^{K}C_K} x_t - (N + 1) = 0 . \tag{3.9}$$

Then

$$\begin{aligned} \underline{a}(\underline{x}) &\equiv \nabla c(\underline{x}) \\ &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} . \end{aligned}$$

Let

$$L(\underline{x}) = f(\underline{x}) + \lambda c(\underline{x}) ,$$

where $\lambda \geq 0$ is a real scalar. Using the method of Lagrange multipliers, those points at which the extrema for this problem occur are amongst the set of solutions to the system

$$\begin{aligned} \nabla L(\underline{x}) = \underline{g}(\underline{x}) + \lambda \underline{a}(\underline{x}) &= \underline{0} \\ c(\underline{x}) &= 0 . \end{aligned}$$

Hence, at an extremum,

$$\psi(x_t + c) - \psi(x_t + 1) + \lambda = 0, \qquad t = 1, \ldots, {}^{r+K}C_K,$$

or

$$-\psi(x_t+c)+\psi(x_t+1)=\lambda, \qquad t=1,\ldots,{}^{r+K}C_K\,. \tag{3.10}$$

Now

$$\psi'(z)=\sum_{n=0}^{\infty}\frac{1}{(z+n)^2}\,,$$

(see [4]) so that

$$\begin{aligned} & -\psi'(x_t+c)+\psi'(x_t+1) \\ = & \left(\frac{-1}{(x_t+c)^2}+\frac{1}{(x_t+1)^2}\right)+\left(\frac{-1}{(x_t+c+1)^2}+\frac{1}{(x_t+2)^2}\right)+ \\ & \cdots+\left(\frac{-1}{(x_t+c+n-1)^2}+\frac{1}{(x_t+n)^2}\right)+\cdots \\ < & \ 0\,. \end{aligned} \tag{3.11}$$

Thus, $-\psi(x_t+c)+\psi(x_t+1)$ is a monotone decreasing function of $x_t$. It follows that the *unique* solution to (3.10) is given by

$$x_1=x_2=\cdots=x_{r+K}{}_{C_K}\,,$$

whence (3.9) gives

$$x_1=x_2=\cdots=x_{r+K}{}_{C_K}=\frac{N+1}{{}^{r+K}C_K}\,. \tag{3.12}$$

Consider the diagonal Hessian matrix

$$W=\begin{bmatrix} \psi'(x_1+c)-\psi'(x_1+1) & & \\ & \ddots & \\ & & \psi'(x_{r+K}{}_{C_K}+c)-\psi'(x_{r+K}{}_{C_K}+1) \end{bmatrix}.$$

At the point given by (3.12),

$$W=\begin{bmatrix} \psi'\left(\frac{N+1}{{}^{r+K}C_K}+c\right)-\psi'\left(\frac{N+1}{{}^{r+K}C_K}+1\right) & & \\ & \ddots & \\ & & \psi'\left(\frac{N+1}{{}^{r+K}C_K}+c\right)-\psi'\left(\frac{N+1}{{}^{r+K}C_K}+1\right) \end{bmatrix}$$

is a positive definite matrix (see (3.11)). Hence, by Theorem 9.3.2 of [30], $f$ has a *local minimum* at this point. However, the fact that there are no other points of extrema implies that $f$ has its *global* minimum here, subject to the constraints.

For $c>1$, a similar argument with $f$ replaced by $-f$ gives the required result. □

**Lemma 3.2.8** *For large $N$,*

$$\sum{}^{+}\left[\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right]\geq\sim N^{({}^{r+K}C_K)c-1},$$

*where $0<c<1, c\in\mathbb{R}$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1})\in\mathcal{H}_{SP}$.*

PROOF: The number of strictly positive histograms based on all $N+1$ groups is the number of ways of putting $N+1-{}^{r+K}C_K$ things into ${}^{r+K}C_K$ boxes, namely ${}^{N}C_{({}^{r+K}C_K-1)}$. Hence, by Lemma 3.2.7,

$$\begin{aligned}\sum{}^{+}\left[\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right] &\geq {}^{N}C_{({}^{r+K}C_K-1)}\left(\frac{\Gamma\left(\frac{N+1}{{}^{r+K}C_K}+c\right)}{\Gamma\left(\frac{N+1}{{}^{r+K}C_K}+1\right)}\right)^{{}^{r+K}C_K}\\ &\sim N^{{}^{r+K}C_K-1}(N^{c-1})^{{}^{r+K}C_K},\quad \text{for large } N,\\ &= N^{({}^{r+K}C_K)c-1}\end{aligned}$$

by Sterling's approximation. □

**Notation**: Throughout the remainder of this chapter, the expression 'const.' refers to a strictly positive, real-valued constant.

Put

$$G(x)=\frac{\Gamma(x+c)}{\Gamma(x+1)},\qquad \forall x\geq 1\,.$$

Then $G$ is decreasing on $[1,\infty)$, for $0<c<1, c\in\mathbb{R}$.

**Lemma 3.2.9** *If $n,p,q\in\mathbb{N}$ and $0<c<1, c\in\mathbb{R}$, then*

$$\sum_{i=1}^{q}G(i)(q+p-i)^{nc-1}\leq const.(q+p)^{(n+1)c-1}\,.$$

PROOF: If $q-1<4$, then

$$\sum_{i=1}^{q}G(i)(q+p-i)^{nc-1}=\text{const.}>0\,,$$

so the result is obvious. Suppose that $q-1\geq 4$, and note that because

$$\frac{G(x)}{x^{c-1}}\to\begin{cases}1, & \text{as } x\to\infty\\ \Gamma(c+1), & \text{as } x\to 1^{+}\end{cases}$$

and the quotient is continuous, it is bounded and hence

$$G(x)\leq \text{const.}\,x^{c-1},\qquad \forall x\geq 1\,.$$

Let

$$f_n(x) = x^{c-1}(q+p-x)^{nc-1}, \qquad \forall x \in [1, q+p-1].$$

Then

$$\begin{aligned} x \leq \frac{q+p}{2} &\Leftrightarrow q+p-x \geq x \\ &\Leftrightarrow \left(\frac{q+p-x}{x}\right)^{(n-1)c} \geq 1 \\ &\Leftrightarrow x^{c-1}(q+p-x)^{nc-1} \geq x^{nc-1}(q+p-x)^{c-1}, \end{aligned}$$

so that

$$f_n(x) \geq f_n(q+p-x), \qquad \forall x \in \left[1, \tfrac{q+p}{2}\right].$$

Hence, if $f_n$ has no critical points in $[1, q+p-1]$, then $f_n$ must be decreasing over this subdomain. Otherwise, elementary calculus shows that $f_n$ is decreasing on $[1, \beta_n]$ where $\beta_n \geq (q+p)/2$. In either case, $f_n$ is decreasing on $\left[1, \frac{q+p}{2}\right]$. Therefore,

$$\begin{aligned} \sum_{i=1}^{q} G(i)(q+p-i)^{nc-1} &\leq \text{const.} \sum_{i=1}^{q} i^{c-1}(q+p-i)^{nc-1} \\ &\leq \text{const.}\, 2 \sum_{i=1}^{\left[\frac{q+p}{2}\right]} i^{c-1}(q+p-i)^{nc-1} \\ &\leq \text{const.} \left((q+p-1)^{nc-1} + \int_1^{\left[\frac{q+p}{2}\right]} x^{c-1}(q+p-x)^{nc-1}dx\right) \\ &\leq \text{const.} \left((q+p-1)^{nc-1} + \int_0^{\frac{q+p}{2}} x^{c-1}(q+p-x)^{nc-1}dx\right). \end{aligned}$$

If $nc-1 \geq 0$, then

$$\begin{aligned} \int_0^{\frac{q+p}{2}} x^{c-1}(q+p-x)^{nc-1}dx &\leq (q+p)^{nc-1} \int_0^{\frac{q+p}{2}} x^{c-1}dx \\ &\leq \text{const.}(q+p)^{(n+1)c-1}. \end{aligned}$$

Otherwise, if $nc-1 < 0$, then

$$\begin{aligned} \int_0^{\frac{q+p}{2}} x^{c-1}(q+p-x)^{nc-1}dx &\leq \left(\frac{q+p}{2}\right)^{nc-1} \int_0^{\frac{q+p}{2}} x^{c-1}dx \\ &\leq \text{const.}(q+p)^{(n+1)c-1}. \end{aligned}$$

Hence, in either case,

$$\begin{aligned}\sum_{i=1}^{q} G(i)(q+p-i)^{nc-1} &\leq \text{const.}\left((q+p-1)^{nc-1} + \text{const.}(q+p)^{(n+1)c-1}\right)\\ &\leq \text{const.}(q+p)^{(n+1)c-1}\,.\end{aligned}$$

□

**Lemma 3.2.10** *For large $N$,*

$$\sum\nolimits^{+}\left[\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right] \leq\sim N^{({}^{r+K}C_K)c-1}\,,$$

*where $0<c<1, c\in\mathbb{R}$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1})\in\mathcal{H}_{SP}$.*

PROOF: Let $R = {}^{r+K}C_K$. Now

$$\sum\nolimits^{+}\left[\prod_{t=1}^{R}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right]$$

may be written

$$\sum_{x_1=1}^{N+2-R}\sum_{x_2=1}^{N+3-R-x_1}\cdots\sum_{x_{R-1}=1}^{N-\sum_{s=1}^{R-2}x_s}\left[\prod_{t=1}^{R-1}\left[\frac{\Gamma(x_t+c)}{\Gamma(x_t+1)}\right]\frac{\Gamma\left(N+1-\sum_{s=1}^{R-1}x_s+c\right)}{\Gamma\left(N+2-\sum_{s=1}^{R-1}x_s\right)}\right].$$

Then recursive use of Lemma 3.2.9 shows that

$$\begin{aligned}&\sum_{x_1=1}^{N+2-R}\sum_{x_2=1}^{N+3-R-x_1}\cdots\sum_{x_{R-1}=1}^{N-\sum_{s=1}^{R-2}x_s}\left[\prod_{t=1}^{R-1}\left[\frac{\Gamma(x_t+c)}{\Gamma(x_t+1)}\right]\frac{\Gamma\left(N+1-\sum_{s=1}^{R-1}x_s+c\right)}{\Gamma\left(N+2-\sum_{s=1}^{R-1}x_s\right)}\right]\\ =&\sum_{x_1=1}^{N+2-R}\sum_{x_2=1}^{N+3-R-x_1}\cdots\sum_{x_{R-1}=1}^{N-\sum_{s=1}^{R-2}x_s}\left[\prod_{t=1}^{R-1}[G(x_t)]G\left(N+1-\sum_{s=1}^{R-1}x_s\right)\right]\\ =&\sum_{x_1=1}^{N+2-R}G(x_1)\sum_{x_2=1}^{N+3-R-x_1}G(x_2)\cdots\sum_{x_{R-1}=1}^{N-\sum_{s=1}^{R-2}x_s}G(x_{R-1})G\left(N+1-\sum_{s=1}^{R-1}x_s\right)\\ \leq&\ \text{const.}\sum_{x_1=1}^{N+2-R}G(x_1)\sum_{x_2=1}^{N+3-R-x_1}G(x_2)\cdots\sum_{x_{R-1}=1}^{N-\sum_{s=1}^{R-2}x_s}G(x_{R-1})\left(N+1-\sum_{s=1}^{R-1}x_s\right)^{c-1}\\ \leq&\ \text{const.}\sum_{x_1=1}^{N+2-R}G(x_1)\sum_{x_2=1}^{N+3-R-x_1}G(x_2)\times\end{aligned}$$

$$\cdots \times \sum_{x_{R-2}=1}^{N+1-\sum_{s=1}^{R-3} x_s} G(x_{R-2}) \left( N+1-\sum_{s=1}^{R-2} x_s \right)^{2c-1}$$

$$\vdots$$

$$\leq \text{ const.} \sum_{x_1=1}^{N+2-R} G(x_1)(N+1-x_1)^{(R-1)c-1}$$

$$\leq \text{ const.}(N+1)^{Rc-1}$$

$$\sim N^{Rc-1}, \quad \text{for large } N\,.$$

□

**Lemma 3.2.11** *For large $N$,*

$$\sum\nolimits^{+} \left[ \prod_{t=1}^{r+K C_K} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \sim N^{(r+K C_K)c-1}\,,$$

*where $c=1$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1}) \in \mathcal{H}_{SP}$.*

PROOF: The number of strictly positive histograms based on all $N+1$ groups is ${}^{N}C_{(r+K C_K-1)}$. Hence,

$$\begin{aligned} \sum\nolimits^{+} \left[ \prod_{t=1}^{r+K C_K} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] &= {}^{N}C_{(r+K C_K-1)} \\ &\sim N^{r+K C_K-1}, \quad \text{for large } N, \\ &= N^{(r+K C_K)c-1} \end{aligned}$$

by Sterling's approximation. □

**Lemma 3.2.12** *For large $N$,*

$$\sum\nolimits^{+} \left[ \prod_{t=1}^{r+K C_K} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \leq\sim N^{(r+K C_K)c-1}\,,$$

*where $c>1, c \in \mathbb{R}$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1}) \in \mathcal{H}_{SP}$.*

PROOF: Follow the proof of Lemma 3.2.8 but reverse the inequality. □

Now put

$$G(x) = \frac{\Gamma(x+c+1)}{\Gamma(x+2)}, \qquad \forall x \geq 0\,.$$

Then $G$ is increasing on $[0,\infty)$, for $c>1, c \in \mathbb{R}$.

**Lemma 3.2.13** *If $n \in \mathbb{N}$, $q \geq p \geq 0$, $p, q \in \mathbb{Z}$ and $c > 1, c \in \mathbb{R}$, then*

$$\sum_{i=p}^{q} G(i-p)(q-i)^{nc-1} \geq const.(q-p)^{(n+1)c-1}\,.$$

PROOF: If $q - p < 4$, then

$$\sum_{i=p}^{q} G(i-p)(q-i)^{nc-1} = \text{const.} \geq 0\,,$$

so the result is obvious. Suppose that $q - p \geq 4$, and note that because

$$\frac{x^{c-1}}{G(x)} \to \begin{cases} 1, & \text{as } x \to \infty \\ 0, & \text{as } x \to 0^+ \end{cases}$$

and the quotient is continuous, it is bounded and hence

$$G(x) \geq \text{const. } x^{c-1}, \qquad \forall x \geq 0\,.$$

Let

$$f_n(x) = (x-p)^{c-1}(q-x)^{nc-1}, \qquad \forall x \in [p, q]\,.$$

Then elementary calculus shows that $f$ is decreasing on $[q - \beta_n, q]$, where

$$\beta_n = \frac{(q-p)(nc-1)}{(nc-1)+(c-1)} \geq \frac{4(nc-1)}{(nc-1)+(c-1)} \geq 2\,.$$

Therefore,

$$\begin{aligned}
\sum_{i=p}^{q} G(i-p)(q-i)^{nc-1} &\geq \text{const.} \sum_{i=p}^{q} (i-p)^{c-1}(q-i)^{nc-1} \\
&\geq \text{const.} \sum_{i=q-[\beta_n]}^{q} (i-p)^{c-1}(q-i)^{nc-1} \\
&\geq \text{const.} \int_{q-[\beta_n]}^{q} (x-p)^{c-1}(q-x)^{nc-1}\,dx \\
&\geq \text{const.}(q-p-[\beta_n])^{c-1} \int_{q-[\beta_n]}^{q} (q-x)^{nc-1}\,dx \\
&\geq \text{const.}(q-p-\beta_n)^{c-1} \int_{q-[\beta_n]}^{q} (q-x)^{nc-1}\,dx \\
&= \text{const.}(q-p)^{c-1}[\beta_n]^{nc} \\
&\geq \text{const.}(q-p)^{c-1} 2^{-nc} (2(\beta_n - 1))^{nc} \\
&\geq \text{const.}(q-p)^{c-1} \beta_n^{nc} \\
&= \text{const.}(q-p)^{(n+1)c-1}\,,
\end{aligned}$$

where $[z]$ denotes the integer part of $z$. □

**Lemma 3.2.14** *For large $N$,*

$$\sum\nolimits^{+}\left[\prod_{t=1}^{{}^{r+K}C_K}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right]\geq\sim N^{({}^{r+K}C_K)c-1},$$

*where $c>1, c\in\mathbb{R}$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1})\in\mathcal{H}_{SP}$.*

PROOF: Let $R={}^{r+K}C_K$. Now

$$\sum\nolimits^{+}\left[\prod_{t=1}^{R}\frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)}\right]$$

may be written

$$\sum_{e_1=0}^{N+1-R}\sum_{e_2=0}^{N+1-R-e_1}\cdots\sum_{e_{R-1}=0}^{N+1-R-\sum_{s=1}^{R-2}e_s}\left[\prod_{t=1}^{R-1}\left[\frac{\Gamma(e_t+1+c)}{\Gamma(e_t+2)}\right]\times\right.$$

$$\left.\frac{\Gamma\left(N+2-R-\sum_{s=1}^{R-1}e_s+c\right)}{\Gamma\left(N+3-R-\sum_{s=1}^{R-1}e_s\right)}\right],$$

or, by putting $S_0=0$, $S_t=\sum_{s=1}^{t}e_s$, $t=1,\ldots,R-1$,

$$\sum_{S_1=0}^{N+1-R}\sum_{S_2=S_1}^{N+1-R}\cdots\sum_{S_{R-1}=S_{R-2}}^{N+1-R}\left[\prod_{t=1}^{R-1}\left[\frac{\Gamma(S_t-S_{t-1}+1+c)}{\Gamma(S_t-S_{t-1}+2)}\right]\frac{\Gamma(N+2-R-S_{R-1}+c)}{\Gamma(N+3-R-S_{R-1})}\right].$$

Then recursive use of Lemma 3.2.13 shows that

$$\begin{aligned}
&\sum_{S_1=0}^{N+1-R}\sum_{S_2=S_1}^{N+1-R}\cdots\sum_{S_{R-1}=S_{R-2}}^{N+1-R}\left[\prod_{t=1}^{R-1}\left[\frac{\Gamma(S_t-S_{t-1}+1+c)}{\Gamma(S_t-S_{t-1}+2)}\right]\times\right.\\
&\qquad\left.\frac{\Gamma(N+2-R-S_{R-1}+c)}{\Gamma(N+3-R-S_{R-1})}\right]\\
=\;&\sum_{S_1=0}^{N+1-R}\sum_{S_2=S_1}^{N+1-R}\cdots\sum_{S_{R-1}=S_{R-2}}^{N+1-R}\left[\prod_{t=1}^{R-1}[G(S_t-S_{t-1})]G(N+1-R-S_{R-1})\right]\\
=\;&\sum_{S_1=0}^{N+1-R}G(S_1)\sum_{S_2=S_1}^{N+1-R}G(S_2-S_1)\times\\
&\qquad\cdots\times\sum_{S_{R-1}=S_{R-2}}^{N+1-R}G(S_{R-1}-S_{R-2})G(N+1-R-S_{R-1})\\
\geq\;&\text{const.}\sum_{S_1=0}^{N+1-R}G(S_1)\sum_{S_2=S_1}^{N+1-R}G(S_2-S_1)\times
\end{aligned}$$

$$\cdots \times \sum_{S_{R-1}=S_{R-2}}^{N+1-R} G(S_{R-1}-S_{R-2})(N+1-R-S_{R-1})^{c-1}$$

$$\geq \text{ const.} \sum_{S_1=0}^{N+1-R} G(S_1) \sum_{S_2=S_1}^{N+1-R} G(S_2-S_1) \times$$

$$\cdots \times \sum_{S_{R-2}=S_{R-3}}^{N+1-R} G(S_{R-2}-S_{R-3})(N+1-R-S_{R-2})^{2c-1}$$

$$\vdots$$

$$\geq \text{ const.} \sum_{S_1=0}^{N+1-R} G(S_1)(N+1-R-S_1)^{(R-1)c-1}$$

$$\geq \text{ const.}(N+1-R)^{Rc-1}$$

$$\sim N^{Rc-1}, \quad \text{for large } N.$$

□

**Lemma 3.2.15** *For large $N$,*

$$\sum\nolimits^{+} \left[ \prod_{t=1}^{r+K C_K} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \sim N^{(r+K C_K)c-1},$$

*where $c>0, c \in \mathbb{R}$ and the summation $\sum^{+}$ runs over all $H(\underline{Y}_{N+1}) \in \mathcal{H}_{SP}$.*

PROOF: The result follows from Lemmas 3.2.8, 3.2.10–3.2.12 and 3.2.14. □

In light of Lemma 3.2.15, we can now state the following.

**Lemma 3.2.16** *For large $N$,*

$$p^* \sim N^{-(r+K C_K-1)c},$$

*where $c>0, c \in \mathbb{R}$ and $p^* = P\left(N+2-{}^{r+K}C_K, 1, \ldots, 1\right)$.*

PROOF: From Equation (3.8),

$$\begin{aligned}
& p^* \\
= & \left(1-\sum\nolimits^{-} P(H(\underline{Y}_{N+1}))\right) \times \\
& \left( \frac{\Gamma\left(N+3-{}^{r+K}C_K\right)}{\Gamma(c+1)^{(r+K C_K-1)}\Gamma\left(N+2-{}^{r+K}C_K+c\right)} \sum\nolimits^{+} \left[ \prod_{t=1}^{r+K C_K} \frac{\Gamma(h_t(N+1)+c)}{\Gamma(h_t(N+1)+1)} \right] \right)^{-1} \\
\sim & \left(N^{1-c} N^{(r+K C_K)c-1}\right)^{-1}, \quad \text{for large } N, \\
= & N^{-(r+K C_K-1)c}
\end{aligned}$$

by Sterling's approximation and Lemma 3.2.15. □

Finally, we are in a position to study the behaviour of (3.7) for large $N$.

**Theorem 3.2.17** *If you assert your predictive probabilities using (3.5) for all $H \in \mathcal{H}_{SP}$, then*

$$\lim_{N\to\infty} P(H(\underline{Y}_M) = h(M)) \neq 0\,,$$

*where $M \leq N$ and $h(M)$ is a strictly positive histogram.*

PROOF: Let $R = {}^{r+K}C_K$. From (3.7),

$$\begin{aligned} P(H(\underline{Y}_M) = h(M)) &= \frac{\Gamma(N+3-R)\Gamma(N+2+Rc-R)}{\Gamma(c+1)^{R-1}\Gamma(N+2)\Gamma(N+2-R+c)} {}^{M}C_{h(M)}\, p^* \times \\ &\qquad \sum{}^{*}\left[ {}^{N+1-M}C_{h(N+1)-h(M)} \frac{\prod_{t=1}^{R}\Gamma(h_t(N+1)+c)}{\Gamma(N+2+Rc-R)} \right] \\ &\sim N^{(R-1)(c-1)} N^{-(R-1)c} N^{R-1}, \quad \text{for large } N, \\ &= 1 \end{aligned}$$

by Sterling's approximation, Lemma 3.2.6 and Lemma 3.2.16. □

Theorem 3.2.17 implies that the systematic use of (3.5) for any size of $N$ as a strategy to estimate predictive probabilities (given any strictly positive histogram) is coherent *without* the need for the concomitant assertion

$$\sum_{H\in\mathcal{H}_{SP}} P(H(\underline{Y}_N) = H) = 0$$

for every value of $N$. Thus, (3.5), and presumably (3.6), appears to represent an improvement to the straight-out frequency mimicking approach.

We will now turn to the study of more sophisticated statistical strategies for specifying predictive probabilities.

# Chapter 4

# Empirical Bayes Estimates

This chapter will investigate the coherency properties of empirical Bayes estimates of predictive probabilities interpreting them as conditional probabilities given the data used to 'estimate' them. In most of the chapter the language used is that of the empirical Bayes setting.

The empirical Bayes approach to the estimation of predictive probabilities is based upon the specification of probability distributions that describe the quantities involved in the problem. A natural assumption, therefore, would be that

$$\begin{aligned} f\left(\underline{Y}^{(i)} \middle| \underline{\theta}^{(i)}\right) &= \frac{r!}{\prod_{j=1}^{K+1} Y_j^{(i)}!} \prod_{j=1}^{K+1} \left[\theta_j^{(i)}\right]^{Y_j^{(i)}} \\ &\sim \text{Multinomial}\left(r, \underline{\theta}^{(i)}\right), \qquad i = 1, \ldots, N+1, \end{aligned}$$

where, in the subjectivist framework, $\underline{\theta}^{(i)}$, $i = 1, \ldots, N+1$, represents a theoretically observable (but presumably unknown) vector of quantities that sum to 1, such as the vector of proportions of the *total* number of items from a very large number of items in the $i^{\text{th}}$ group that would be classified into each of the $K+1$ categories. The question of what type of distribution to put on $\underline{\theta}^{(i)}$, $i = 1, \ldots, N+1$, is one that has concerned statisticians for some time. The natural conjugate distribution to the Multinomial is the Dirichlet, given by

$$\begin{aligned} f\left(\underline{\theta}^{(i)} \middle| \underline{\alpha}\right) &= \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \prod_{j=1}^{K+1} \left[\theta_j^{(i)}\right]^{\alpha_j - 1} \\ &\sim \text{Dirichlet}(\underline{\alpha}), \qquad i = 1, \ldots, N+1, \end{aligned} \tag{4.1}$$

where $\underline{\alpha} > \underline{0}$. The $\underline{\theta}^{(i)}$ vectors, $i = 1, \ldots, N+1$, are presumed independent conditional upon $\underline{\alpha}$, a structure the subjectivist would recognise as motivating an exchangeable distribution over the $\underline{\theta}^{(i)}$. Some of the nice features of the Dirichlet distribution are that it is mathematically tractable and provides a large class of distributions on the simplex. However, if $\underline{\theta}$ has a Dirichlet($\underline{\alpha}$) distribution, then the covariance of any two components of $\underline{\theta}$ is negative. That is,

$$\mathrm{Cov}[\theta_j, \theta_k] < 0, \qquad j \neq k\,.$$

This may be undesirable if, for instance, there is believed to be an ordering present amongst the categories, for then it might be expected that $\underline{\theta}$ components corresponding to neighbouring categories have a positive covariance. This reason has prompted a number of authors to search for alternatives to the Dirichlet distribution to describe vectors of proportions. Connor and Mosimann [19] introduce the concept of 'neutrality' for such vectors and derive a generalisation of the Dirichlet distribution, while Dennis [26] generalises this further in producing what he calls a hyper-Dirichlet type 1 distribution. Aitchison and Shen [1] summarise the properties and uses of the logistic-Normal distribution and Aitchison [2] introduces a general class of distributions on the simplex which includes as special cases the Dirichlet and logistic-Normal classes. Aitchison [3] gives a thorough review of the analysis of compositional data. Dickey [27] develops a class of distributions that amounts to a distribution over convex combinations of Dirichlet distributed vectors. The analysis in this chapter can be extended to this context, but as will be seen, the detail is already quite complex. Pursuing it in the setting of a larger family of distributions would detract from the clarity of the logic of the approach to be developed. Hence (4.1) will be adopted throughout the remainder of this chapter.

If the parameter $\underline{\alpha}$ were *known*, knowledge of the outcomes of the first $N$ groups would be redundant. Your predictive probabilities would therefore be calculated using

$$\begin{aligned}
& f\left(\underline{Y}^{(N+1)} \middle| \underline{\alpha}\right) \\
&= \int_{\underline{\theta}^{(N+1)}} \frac{r!}{\prod_{j=1}^{K+1} Y_j^{(N+1)}!} \prod_{j=1}^{K+1} \left[\theta_j^{(N+1)}\right]^{Y_j^{(N+1)}} \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \prod_{j=1}^{K+1} \left[\theta_j^{(i)}\right]^{\alpha_j - 1} d\underline{\theta}^{(N+1)} \\
&= \frac{r!}{\prod_{j=1}^{K+1} Y_j^{(N+1)}!} \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \int_{\underline{\theta}^{(N+1)}} \prod_{j=1}^{K+1} \left[\theta_j^{(N+1)}\right]^{Y_j^{(N+1)} + \alpha_j - 1} d\underline{\theta}^{(N+1)}
\end{aligned}$$

$$= \frac{r!}{\prod_j^{K+1} Y_j^{(N+1)}!} \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \frac{\prod_{j=1}^{K+1} \Gamma\left(\alpha_j + Y_j^{(N+1)}\right)}{_{=1}\ \Gamma\left(\sum_{j=1}^{K+1} \alpha_j + r\right)}, \tag{4.2}$$

which is the marginal density of $\underline{Y}^{(N+1)}$. The distribution given by (4.2) is commonly referred to as the Dirichlet-Multinomial distribution and denoted DMD$(r, \underline{\alpha})$. It has also been called the compound Multinomial distribution, the multivariate Pólya-Eggenberger distribution and the multivariate or multiple category Pólya distribution (due to its known equivalence with Pólya urn models) by other authors. (See [51] for a review of different ways in which the Dirichlet-Multinomial distribution can be derived.) Including the Beta-Binomial ($K = 1$) as a special case, the Dirichlet-Multinomial distribution has been widely used in the social, physical and health sciences to describe the extra variability in observed counts which cannot be explained by a Binomial or Multinomial model. Applications have been made to fossil pollen data (Mosimann [72]), consumer purchasing behaviour (Chatfield and Goodhardt [16], Goodhardt *et al.* [45], Lenk [65]), epidemiology (Griffiths [46]), teratological data from litters of animals (Williams [90], Haseman and Kupper [47], Segreti and Munson [82]), magazine and television advertising exposure (Chandon [15], Leckenby and Kishi [63], Rust and Leone [80], Danaher [21]), data on the differential blood count (Unkelbach [86]) and sequential bidding for contracts (Attwell and Smith [6]). In the context of cluster sampling, the distribution has been used to model dependence between observations within the same cluster (Brier [12], Wilson [91], Koehler and Wilson [55]). The Dirichlet-Multinomial distribution has also been employed in discussions of the Bayes and pseudo-Bayes analysis of categorical data (Good [38, 40], Hoadley [50], Good and Crook [44], Bishop *et al.* [8], Lee and Sabavala [64]).

Of course the parameter $\underline{\alpha}$ is usually *unknown* in most practical situations and is dealt with in one of two ways. Either $\underline{\alpha}$ is estimated from the data already available or a distribution is placed on $\underline{\alpha}$. The former is the approach taken by empirical Bayes statisticians and is the subject of this chapter, while the latter hierarchical Bayesian approach is considered in Chapter 5.

Once an estimate, $\hat{\underline{\alpha}}$, of $\underline{\alpha}$ has been obtained from the outcomes of the first $N$

groups it will be substituted into (4.2) to give

$$P\left(\underline{Y}^{(N+1)} = \underline{y} \,\middle|\, \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) = \frac{r!\Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j\right) \prod_{j=1}^{K+1} \Gamma\left(\hat{\alpha}_j + y_j\right)}{\prod_{j=1}^{K+1}[y_j!] \prod_{j=1}^{K+1}[\Gamma(\hat{\alpha}_j)]\Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j + r\right)}. \quad (4.3)$$

It remains to study some of the many suggested methods for estimating $\underline{\alpha}$ and their effects on the coherency of the resulting predictive probabilities.

## 4.1 Method of Moments Estimates

The moment approach to estimating $\underline{\alpha}$ applies when it is possible to relate prior moments to moments of the marginal distribution, the latter being supposedly estimated from the data or determined subjectively.

If

$$\begin{aligned} f(\underline{Y} \,|\, \underline{\theta}) &\sim \text{Multinomial}(r, \theta_1, \ldots, \theta_{K+1}) \\ f(\underline{\theta} \,|\, \underline{\alpha}) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{K+1}), \end{aligned}$$

then

$$f(\underline{Y} \,|\, \underline{\alpha}) \sim \text{DMD}(r, \alpha_1, \ldots, \alpha_{K+1}).$$

Hence, using the moments of the Dirichlet distribution (see Wilks [89]),

$$\begin{aligned} E[Y_j] &= E[E[Y_j \,|\, \underline{\theta}]] \\ &= E[r\theta_j] \\ &= \frac{r\alpha_j}{\sum_{l=1}^{K+1} \alpha_l}, \end{aligned} \quad (4.4)$$

$$\begin{aligned} \text{Var}[Y_j] &= \text{Var}[E[Y_j \,|\, \underline{\theta}]] + E[\text{Var}[Y_j \,|\, \underline{\theta}]] \\ &= \text{Var}[r\theta_j] + E[r\theta_j(1-\theta_j)] \\ &= \frac{r^2\alpha_j\left(\sum_{l=1}^{K+1} \alpha_l - \alpha_j\right)}{\left(\sum_{l=1}^{K+1} \alpha_l + 1\right)\left(\sum_{l=1}^{K+1} \alpha_l\right)^2} + \left(\frac{r\alpha_j}{\sum_{l=1}^{K+1} \alpha_l} - \frac{r(\alpha_j+1)\alpha_j}{\left(\sum_{l=1}^{K+1} \alpha_l + 1\right)\sum_{l=1}^{K+1} \alpha_l}\right) \\ &= \frac{r\alpha_j\left(\sum_{l=1}^{K+1} \alpha_l - \alpha_j\right)\left(r + \sum_{l=1}^{K+1} \alpha_l\right)}{\left(\sum_{l=1}^{K+1} \alpha_l\right)^2\left(1 + \sum_{l=1}^{K+1} \alpha_l\right)} \\ &= rE[\theta_j](1 - E[\theta_j])\left(\frac{r + \sum_{l=1}^{K+1} \alpha_l}{1 + \sum_{l=1}^{K+1} \alpha_l}\right) \end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}[Y_j Y_k] &= E[Y_j Y_k] - E[Y_j]E[Y_k] \\
&= E[E[Y_j Y_k \mid \underline{\theta}]] - E[Y_j]E[Y_k] \\
&= E[r(r-1)\theta_j\theta_k] - E[Y_j]E[Y_k] \\
&= \frac{r(r-1)\alpha_j\alpha_k}{\left(\sum_{l=1}^{K+1}\alpha_l + 1\right)\sum_{l=1}^{K+1}\alpha_l} - \frac{r^2\alpha_j\alpha_k}{\left(\sum_{l=1}^{K+1}\alpha_l\right)^2} \\
&= \frac{-r\alpha_j\alpha_k\left(r + \sum_{l=1}^{K+1}\alpha_l\right)}{\left(\sum_{l=1}^{K+1}\alpha_l\right)^2\left(1 + \sum_{l=1}^{K+1}\alpha_l\right)} \\
&= -rE[\theta_j]E[\theta_k]\left(\frac{r + \sum_{l=1}^{K+1}\alpha_l}{1 + \sum_{l=1}^{K+1}\alpha_l}\right).
\end{aligned}$$

Noting that

$$\mathrm{Var}[Y_j \mid \underline{\theta}] = r\theta_j(1-\theta_j)$$

and

$$\mathrm{Cov}[Y_j Y_k \mid \underline{\theta}] = -r\theta_j\theta_k\,,$$

we follow Mosimann [72] in observing the result

$$\Sigma_{\mathrm{DMD}} = \left(\frac{r + \sum_{l=1}^{K+1}\alpha_l}{1 + \sum_{l=1}^{K+1}\alpha_l}\right)\Sigma_{\mathrm{Mult}}\,, \tag{4.5}$$

where $\Sigma_{\mathrm{DMD}}$ is the covariance matrix of the Dirichlet-Multinomial distribution and $\Sigma_{\mathrm{Mult}}$ denotes the covariance matrix of the Multinomial distribution were $\underline{\theta}$ fixed and equal to its expected value, $E[\underline{\theta}]$, as a function of $\underline{\alpha}$. The constant multiplier in (4.5) is usually denoted in the literature by

$$C = \frac{r + \sum_{l=1}^{K+1}\alpha_l}{1 + \sum_{l=1}^{K+1}\alpha_l}\,, \tag{4.6}$$

where, clearly, $1 \leq C < r$.

Equation (4.4) suggests estimating the parameters of the Dirichlet-Multinomial distribution by setting each sample mean equal to its expectation

$$\bar{Y}_j = \frac{r\hat{\alpha}_j}{\hat{\tau}}, \qquad j = 1, \ldots, K+1\,, \tag{4.7}$$

where

$$\bar{Y}_j = \frac{1}{N}\sum_{i=1}^{N} Y_j^{(i)}, \qquad \tau = \sum_{l=1}^{K+1}\alpha_l\,.$$

However, there are only $K$ linearly independent equations in (4.7) due to the fact that the components of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, must sum to $r$. Equation (4.6) suggests that an estimate, $\hat{C}$, of $C$ be obtained using (4.5), to provide one further equation

$$\hat{C} = \frac{r + \hat{\tau}}{1 + \hat{\tau}}, \tag{4.8}$$

where $r > 1$ is needed for identifiability of $\tau$.

Thus the estimation of $\underline{\alpha}$ naturally separates into estimation of the quantities $\lambda_1 \equiv \alpha_1/\tau$, ..., $\lambda_K \equiv \alpha_K/\tau$ and $\tau$. For this reason, many authors choose to write the parameters of the Dirichlet distribution as $\tau\lambda_1, \ldots, \tau\lambda_{K+1}$ where $\tau > 0$, $\lambda_j > 0$, $j = 1, \ldots, K+1$, and $\sum_{j=1}^{K+1} \lambda_j = 1$.

In the literature, proponents of the method of moments approach have unanimously adopted (4.7) as the first step in estimating $\underline{\alpha}$. The point of contention in the use of (4.8) is how to best estimate $C$. We begin by studying two of the most commonly quoted estimators.

### 4.1.1 Mosimann's $\hat{C}$

Mosimann [72] suggested estimating $C$ by

$$\hat{C} = \left( \frac{|W_{\text{DMD}}|}{|W_{\text{Mult}}|} \right)^{1/K}, \tag{4.9}$$

where $W_{\text{Mult}}$ and $W_{\text{DMD}}$ are consistent estimates of $\Sigma_{\text{Mult}}$ and $\Sigma_{\text{DMD}}$, respectively. For example, possible terms of $W_{\text{Mult}}$ are

$$\begin{aligned} w_{jj} &= \frac{\bar{Y}_j \left( r - \bar{Y}_j \right)}{r} \\ w_{jk} &= \frac{-\bar{Y}_j \bar{Y}_k}{r} \end{aligned} \tag{4.10}$$

and possible terms of $W_{\text{DMD}}$ are

$$\begin{aligned} w_{jj} &= \frac{\sum_{i=1}^{N} \left( Y_j^{(i)} - \bar{Y}_j \right)^2}{N - 1} \\ w_j &= \frac{\sum_{i=1}^{N} \left[ \left( Y_j^{(i)} - \bar{Y}_j \right) \left( Y_k^{(i)} - \bar{Y}_k \right) \right]}{N - 1} \end{aligned} \tag{4.11}$$

(or replace $N - 1$ by $N$ in both denominators of (4.11)), where $j, k \in \{1, \ldots, K\}$. Note that the ratio of the observed variance of $Y_j^{(i)}$ ($j \in \{1, \ldots, K+1\}$) to its

Multinomial counterpart, for example, could be used to determine $C$, but an estimate based on one variable only is probably less desirable than an estimate using all variances and covariances. In constructing $W_{\text{Mult}}$ and $W_{\text{DMD}}$ it is necessary to consider only $K$ of the $K+1$ categories in order to avoid singularity. The following theorem shows that the choice of which $K$ categories to include has no effect on the resulting value of $\hat{C}$, a question unaddressed in Mosimann's work.

**Theorem 4.1.1** *Let $W_{Mult}$ and $W_{DMD}$ be the $(K+1)\times(K+1)$ matrices whose entries are given by (4.10) and (4.11), respectively. Then*

$$|(W_{Mult})_{11}| = |(W_{Mult})_{22}| = \cdots = |(W_{Mult})_{K+1,K+1}|$$

*and*

$$|(W_{DMD})_{11}| = |(W_{DMD})_{22}| = \cdots = |(W_{DMD})_{K+1,K+1}|,$$

*where $|A_{jj}|$ denotes the determinant of the submatrix obtained by deleting row and column $j$ from $A$.*

PROOF: Let $W = W_{\text{Mult}}$ and fix $k \in \{1,\ldots,K+1\}$. Then

$$\begin{aligned}\sum_{j=1}^{K+1} w_{jk} &= \sum_{\substack{j=1\\ j\neq k}}^{K+1}\left[\frac{-\bar{Y}_j\bar{Y}_k}{r}\right] + \frac{\bar{Y}_k\left(r-\bar{Y}_k\right)}{r}\\ &= \frac{\bar{Y}_k}{r}\left(r-\bar{Y}_k-\sum_{\substack{j=1\\ j\neq k}}^{K+1}\bar{Y}_j\right)\\ &= 0 \qquad (4.12)\end{aligned}$$

and (4.12) holds for all $k \in \{1,\ldots,K+1\}$. Hence, all of the rows and columns of $W$ sum to 0, since $W$ is symmetric. If $K = 1$ the result is obvious, so assume $K > 1$. The theorem has a nice proof using the rule of false cofactors, however a different argument which is somewhat shorter is presented here. Let

$$S = \begin{bmatrix} -1 & -1 & -1 & \cdots & -1\\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

be the $K \times K$ matrix obtained by replacing the first row of the identity matrix by a row of $-1$'s. Let $T_2 = I_K$ and for $k = 3, \ldots, K+1$, let

$$T_k = \begin{bmatrix} 0 & 1 & & & & & & \\ & 0 & 1 & & & & & \\ & & 0 & & & & & \\ & & & \ddots & 1 & & & \\ 1 & & & & 0 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix} \tag{4.13}$$

be the $K \times K$ permutation matrix obtained by interchanging the first $k-1$ rows of the identity matrix in the manner suggested by (4.13). Then,

$$W_{11} = T_k S W_{kk} S^T T_k^T, \qquad k = 2, \ldots, K+1,$$

so that

$$\begin{aligned} |W_{11}| &= |T_k|^2 |S|^2 |W_{kk}| \\ &= (\pm 1)^2 (-1)^2 |W_{kk}| \\ &= |W_{kk}|, \qquad k = 2, \ldots, K+1. \end{aligned}$$

A similar argument applies in the case of $W_{\text{DMD}}$. □

**Example 4.1.1** *To illustrate the calculation of Mosimann's $\hat{C}$, consider the data from Mosimann [72], reproduced in Table B.1 of Appendix B, §B.1. The data are counts of the frequencies of occurrence of different kinds of pollen grains, made at various levels of a core, with the attempt to reconstruct past vegetation changes in the area from which the core was taken. Four arboreal types of pollen were counted until a total of 100 grains was reached. Henceforth the subscripts 1, 2, 3, 4 refer to pine (*Pinus*), fir (*Abies*), oak (*Quercus*) and alder (*Alnus*) pollen, respectively. The data contain 73 sets of counts, each representing a different depth in the core.*

*For this data, $N = 73$, $r = 100$, $K + 1 = 4$ and*

$$\begin{aligned} \bar{Y}_1 &= 6298/73 \\ \bar{Y}_2 &= 103/73 \\ \bar{Y}_3 &= 662/73 \\ \bar{Y}_4 &= 237/73\,. \end{aligned} \tag{4.14}$$

*Also,*

$$\begin{aligned} \sum_{i=1}^{73} \left(Y_1^{(i)} - \bar{Y}_1\right)^2 &= 254954/73 \\ \sum_{i=1}^{73} \left(Y_2^{(i)} - \bar{Y}_2\right)^2 &= 10926/73 \\ \sum_{i=1}^{73} \left(Y_3^{(i)} - \bar{Y}_3\right)^2 &= 135974/73 \\ \sum_{i=1}^{73} \left(Y_4^{(i)} - \bar{Y}_4\right)^2 &= 43038/73 \\ \sum_{i=1}^{73} \left[\left(Y_1^{(i)} - \bar{Y}_1\right)\left(Y_2^{(i)} - \bar{Y}_2\right)\right] &= -15930/73 \\ \sum_{i=1}^{73} \left[\left(Y_1^{(i)} - \bar{Y}_1\right)\left(Y_3^{(i)} - \bar{Y}_3\right)\right] &= -166832/73 \\ \sum_{i=1}^{73} \left[\left(Y_1^{(i)} - \bar{Y}_1\right)\left(Y_4^{(i)} - \bar{Y}_4\right)\right] &= -72192/73 \\ \sum_{i=1}^{73} \left[\left(Y_2^{(i)} - \bar{Y}_2\right)\left(Y_3^{(i)} - \bar{Y}_3\right)\right] &= 3354/73 \\ \sum_{i=1}^{73} \left[\left(Y_2^{(i)} - \bar{Y}_2\right)\left(Y_4^{(i)} - \bar{Y}_4\right)\right] &= 1650/73 \\ \sum_{i=1}^{73} \left[\left(Y_3^{(i)} - \bar{Y}_3\right)\left(Y_4^{(i)} - \bar{Y}_4\right)\right] &= 27504/73\,. \end{aligned} \tag{4.15}$$

*Hence, using categories 2, 3 and 4 in (4.10) and (4.11),*

$$W_{Mult} = \begin{bmatrix} 741291/532900 & -34093/266450 & -24411/532900 \\ -34093/266450 & 1098589/133225 & -78447/266450 \\ -24411/532900 & -78447/266450 & 1673931/532900 \end{bmatrix}$$

*and*

$$W_{DMD} = \begin{bmatrix} 607/292 & 559/876 & 275/876 \\ 559/876 & 67987/2628 & 382/73 \\ 275/876 & 382/73 & 2391/292 \end{bmatrix}.$$

*Then, by (4.9),*

$$\begin{aligned}\hat{C} &= \left(\frac{|W_{DMD}|}{|W_{Mult}|}\right)^{1/3} \\ &= 2.1962 \quad (4 \text{ d.p.}),\end{aligned}$$

*so that, by (4.8),*

$$\begin{aligned}\hat{\tau} &= \frac{100-\hat{C}}{\hat{C}-1} \\ &= 81.76 \quad (2 \text{ d.p.}).\end{aligned}$$

We now turn to studying the coherency of predictive probabilities asserted on the basis of these estimators. We begin with a definition.

**Definition 4.1.1** *A histogram, $H$, is said to be* ***positive*** *if at least one item from the $N$ groups has been observed in each category,* i.e., *if $\bar{Y}_j > 0$, $j = 1, \ldots, K+1$.*

Note that a positive histogram does not necessarily contain a positive number of each type of outcome for $\underline{Y}^{(i)}$. Hence, $\mathcal{H}_P \not\subseteq \mathcal{H}_{SP}$, although $\mathcal{H}_{SP} \subseteq \mathcal{H}_P$.

For (4.7) to give valid parameter estimates requires that a positive histogram has been observed, otherwise some $\hat{\alpha}_j = 0$ which would be improper. Also, it has already been noted that $1 \leq C < r$, for $\tau > 0$. However Mosimann's estimate, $\hat{C}$, using (4.10) and (4.11) is not guaranteed to lie in this interval. Therefore, let $\mathcal{H}_M$ be the set of positive histograms for which Mosimann's $\hat{C}$ (using (4.10) and (4.11)) satisfies $1 < \hat{C} < r$. It then seems sensible to investigate the coherency of your opinions if you would assert

$$p_{t,H} = \frac{r!\Gamma\left(\sum_{j=1}^{K+1}\hat{\alpha}_j\right)\prod_{j=1}^{K+1}\Gamma\left(\hat{\alpha}_j + a_j\right)}{\prod_{j=1}^{K+1}[a_j!]\prod_{j=1}^{K+1}[\Gamma(\hat{\alpha}_j)]\Gamma\left(\sum_{j=1}^{K+1}\hat{\alpha}_j + r\right)}, \qquad t = 1, \ldots, {}^{r+K}C_K\,, \qquad (4.16)$$

given $H \in \mathcal{H}_M$, where type $t$ is

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{K+1} \end{pmatrix}$$

and $\underline{\hat{\alpha}}$ is determined by solving (4.7) and (4.8), using (4.9), (4.10) and (4.11).

Unfortunately, there are two main problems with the use of Theorem 2.4.1 in facilitating a nice theoretical analysis of the coherency of the strategy given by (4.16). Firstly, it is not at all clear, for given values of $N$, $r$ and $K+1$, whether or not the set $\mathcal{H}_M$ is linked. If not, Theorem 2.4.1 still applies to all linked subsets of $\mathcal{H}_M$, but these subsets may be difficult to identify or classify in general. The second problem has to do with looking at $n$-cycle probability ratios, in particular complicated expressions such as $p_{i,h}/p_{j,h}$ where the probabilities are of the form (4.16). Again, it is not obvious in general whether or not these $n$-cycles are satisfied. It was therefore decided to take an empirical approach to the study of (4.16). This was accomplished by developing an algorithm which takes the values $N+1$, $r$ and $K+1$ as input, finds the set of histograms, $\mathcal{H}_M$, and derives the system of coherency induced equations to be solved. The algorithm was implemented in MAPLE V RELEASE 3 (see Appendix A, §A.1) and the SOLVE routine of this package was used to solve the resulting system of equations. Another algorithm was developed, and then implemented in MAPLE V RELEASE 3, to test whether or not a given system of equations, generated by histograms from some set, is linked (see Appendix A, §A.3). The results for problems of various sizes are presented in Table 4.1. Note that due to the necessary trial and error approach to determining whether or not a given histogram is a member of the set $\mathcal{H}_M$, the program is quite time expensive and so only small-sized problems were practical for consideration by this method. The following explanations apply to expressions used in Table 4.1.

- histcount — Number of histograms in $\mathcal{H}_M$.
- dimhist — Number of histograms in $\mathcal{H}'$, *i.e.*, total number of possible histograms.
- qcount — Number of variables, *i.e.*, components of $\underline{q}$, involved in the system of equations generated by histograms from $\mathcal{H}_M$.
- dimq — Total number of variables, *i.e.*, components of $\underline{q}$.
- eqncount — Number of coherency induced equations, including the condition that all of the variables sum to 1 ($\underline{q}^T\underline{1} = 1$).
- $l$ — Number of linked subsets of $\mathcal{H}_M$.

- $q$'s fn $q_*$ — The system of equations generated by histograms from $\mathcal{H}_M$ has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.
- $q$'s $= 0$ — Only the trivial solution to the system of equations generated by histograms from $\mathcal{H}_M$ exists.

| $N+1$ | $r$ | $K+1$ | histcount/ dimhist | qcount/ dimq | eqncount | $l$ | Solution |
|---|---|---|---|---|---|---|---|
| 3 | 2 | 3 | 0/21 | - | - | - | - |
| 4 | 2 | 3 | 9/56 | 42/126 | 46 | 1 | $q$'s fn $q_*$ |
| 5 | 2 | 3 | 30/126 | 102/252 | 151 | 1 | $q$'s $= 0$ |
| 6 | 2 | 3 | 84/252 | 220/462 | 421 | 1 | $q$'s $= 0$ |
| 3 | 3 | 3 | 0/55 | - | - | - | - |
| 4 | 3 | 3 | 59/220 | 369/715 | 532 | 1 | $q$'s $= 0$ |

Table 4.1: Results for Mosimann's $\hat{C}$

The results in Table 4.1 suggest, for given values of $N+1$, $r$ and $K+1$, it is quite likely that the only way for you to be coherent in using the strategy given by (4.16) is to give zero probability to exactly those situations in which it would apply! Furthermore, for fixed $r$ and $K+1$, there may always exist $N^* \geq 2$ such that this is true for all $N \geq N^*$. These issues will be reinvestigated after looking at other possible estimators of $C$.

### 4.1.2 Brier's $\hat{C}$

The analysis of contingency tables under cluster sampling led Brier [11, 12] to suggest estimating $C$ by

$$\hat{C} = \frac{1}{(N-1)K} \sum_{i=1}^{N} \sum_{j=1}^{K+1} \frac{\left(Y_j^{(i)} - \bar{Y}_j\right)^2}{\bar{Y}_j}. \tag{4.17}$$

This is a consistent estimator of $C$ and is intuitively appealing as it is the Pearson chi-squared statistic for testing equality of the vectors $\underline{\theta}^{(i)}$, $i = 1, \ldots, N$, divided by the corresponding degrees of freedom.

**Example 4.1.2** *To illustrate the calculation of Brier's $\hat{C}$, consider again the data from Mosimann [72], reproduced in Table B.1 of Appendix B, §B.1. For this data, recall $N = 73$, $r = 100$, $K+1 = 4$ and (4.14), (4.15) from Example 4.1.1. Then, by (4.17),*

$$\begin{aligned}\hat{C} &= \frac{1}{72 \times 3} \sum_{i=1}^{73} \sum_{j=1}^{4} \frac{\left(Y_j^{(i)} - \bar{Y}_j\right)^2}{\bar{Y}_j} \\ &= \frac{1}{72 \times 3} \left( \frac{254954}{6298} + \frac{10926}{103} + \frac{135974}{662} + \frac{43038}{237} \right) \\ &= 2.4702 \quad (4 \text{ d.p.}),\end{aligned}$$

*so that, by (4.8),*

$$\begin{aligned}\hat{\tau} &= \frac{100 - \hat{C}}{\hat{C} - 1} \\ &= 66.34 \quad (2 \text{ d.p.}). \end{aligned} \tag{4.18}$$

It is worth mentioning here that Brier's estimate of $\tau$ for this data set appears to have been miscalculated, or at least misquoted, in several papers in the literature. Chuang and Cox in Table I of [17] and Danaher in Table III of [21] give Brier's $\hat{\tau} = 73.21$ (2 d.p.) for Mosimann's pollen data, perhaps making it seem less attractive when compared to other estimates than is really the case. Private correspondence with the aforementioned authors has led to both accepting (4.18) as the correct figure.

Clearly, both (4.7) and (4.17) will only give valid parameter estimates if a positive histogram has been observed. Otherwise (4.17) would involve a division by zero, and (4.7) would yield some $\hat{\alpha}_j = 0$ which would be improper. Brier's estimate, $\hat{C}$, also suffers from the problem that it is not guaranteed to lie in the interval $[1, r)$. Brier [11, 12] recommends truncating the value of $\hat{C}$ to be either 1 or $r$ should it fall outside of this interval. This suggestion may be appropriate in situations such as hypothesis testing where an estimate of $C$ is all that is required, however the implied estimates of $\infty$ or 0 for $\tau$ are invalid in the present context. Therefore, we now define $\mathcal{H}_B$ to be the set of positive histograms for which Brier's $\hat{C}$ satisfies $1 < \hat{C} < r$. It then seems sensible to investigate the coherency of your opinions if you would assert

$$p_{t,H} = \frac{r! \Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j\right) \prod_{j=1}^{K+1} \Gamma\left(\hat{\alpha}_j + a_j\right)}{\prod_{j=1}^{K+1} [a_j!] \prod_{j=1}^{K+1} [\Gamma(\hat{\alpha}_j)] \Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j + r\right)}, \qquad t = 1, \ldots, {}^{r+K}C_K, \tag{4.19}$$

given $H \in \mathcal{H}_B$, where type $t$ is

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{K+1} \end{pmatrix}$$

and $\hat{\underline{\alpha}}$ is determined by solving (4.7) and (4.8), using (4.17).

The same problems that arise with analysing the coherency of the strategy given by (4.16) present themselves when considering the strategy given by (4.19): namely, difficulty determining whether or not $\mathcal{H}_B$ is linked and whether or not a general $n$-cycle is satisfied. It was therefore decided to take an empirical approach to the study of (4.19) also. This was accomplished by developing an algorithm which takes the values $N+1$, $r$ and $K+1$ as input, finds the set of histograms, $\mathcal{H}_B$, and derives the system of coherency induced equations to be solved. The algorithm was implemented in MAPLE V RELEASE 3 (see Appendix A, §A.2) and the SOLVE routine of this package was used to solve the resulting system of equations. This system was tested, using the previously developed algorithm (see Appendix A, § A.3), to see whether or not it is linked. The results for problems of various sizes are presented in Table 4.2. Note that due to the necessary trial and error approach to determining whether or not a given histogram is a member of the set $\mathcal{H}_B$, the program is quite time expensive. Again only small-sized problems were practical for consideration by this method, although a larger number of results can be presented due to the form of Brier's $\hat{C}$ being much simpler than Mosimann's. The expressions used in Table 4.2 have the same explanations as before in Table 4.1, with $\mathcal{H}_B$ in place of $\mathcal{H}_M$. Table 4.2 refers to multiple $q_*$'s appropriately only in cases where multiple distinct blocks of linked histograms occur.

The results in Table 4.2 suggest, for given values of $N+1$, $r$ and $K+1$, it is highly likely that the only way for you to be coherent in using the strategy given by (4.19) is to give zero probability to exactly those situations in which it would apply! Furthermore, for fixed $r$ and $K+1$, there may exist $N^* \geq 2$ such that this is true, and $\mathcal{H}_B$ is linked, for all $N \geq N^*$. These issues will be reinvestigated after looking at other possible estimators of $C$.

Many authors have produced estimators of $C$ when looking for correction factors to the standard $\chi^2$ test statistics used in the analysis of categorical data arising

| $N+1$ | $r$ | $K+1$ | histcount/ dimhist | qcount/ dimq | eqncount | $l$ | Solution |
|---|---|---|---|---|---|---|---|
| 3 | 2 | 3 | 0/21 | - | - | - | - |
| 4 | 2 | 3 | 15/56 | 60/126 | 76 | 1 | $q$'s $= 0$ |
| 5 | 2 | 3 | 51/126 | 147/252 | 256 | 1 | $q$'s $= 0$ |
| 6 | 2 | 3 | 123/252 | 295/462 | 616 | 1 | $q$'s $= 0$ |
| 7 | 2 | 3 | 238/462 | 513/792 | 1191 | 1 | $q$'s $= 0$ |
| 3 | 2 | 4 | 3/55 | 30/220 | 28 | 3 | $q$'s fn 3 $q_*$'s |
| 4 | 2 | 4 | 24/220 | 190/715 | 217 | 1 | $q$'s fn $q_*$ |
| 5 | 2 | 4 | 212/715 | 920/2002 | 1909 | 1 | $q$'s $= 0$ |
| 6 | 2 | 4 | 836/2002 | 2787/5005 | 7525 | 1 | $q$'s $= 0$ |
| 3 | 2 | 5 | 0/120 | - | - | - | - |
| 4 | 2 | 5 | 45/680 | 535/3060 | 631 | 1 | $q$'s fn $q_*$ |
| 5 | 2 | 5 | 365/3060 | 2925/11628 | 5111 | 1 | $q$'s $= 0$ |
| 3 | 2 | 6 | 0/231 | - | - | - | - |
| 4 | 2 | 6 | 15/1771 | 315/10626 | 301 | 15 | $q$'s fn 15 $q_*$'s |
| 5 | 2 | 6 | 690/10626 | 8490/53130 | 13801 | 1 | $q$'s $= 0$ |
| 3 | 3 | 3 | 12/55 | 101/220 | 109 | 2 | $q$'s $= 0$, $q$'s fn $q_*$ |
| 4 | 3 | 3 | 113/220 | 534/715 | 1018 | 1 | $q$'s $= 0$ |
| 5 | 3 | 3 | 447/715 | 1597/2002 | 4024 | 1 | $q$'s $= 0$ |
| 6 | 3 | 3 | 1432/2002 | 4200/5005 | 12889 | 1 | $q$'s $= 0$ |
| 3 | 3 | 4 | 40/210 | 644/1540 | 761 | 1 | $q$'s $= 0$ |
| 4 | 3 | 4 | 588/1540 | 5549/8855 | 11173 | 1 | $q$'s $= 0$ |
| 3 | 4 | 3 | 48/120 | 471/680 | 673 | 1 | $q$'s $= 0$ |
| 4 | 4 | 3 | 444/680 | 2607/3060 | 6217 | 1 | $q$'s $= 0$ |
| 3 | 4 | 4 | 181/630 | 4495/7770 | 6155 | 1 | $q$'s $= 0$ |
| 4 | 4 | 4 | 4221/7770 | 57772/73815 | 143515 | 1 | $q$'s $= 0$ |
| 3 | 5 | 3 | 111/231 | 1381/1771 | 2221 | 1 | $q$'s $= 0$ |
| 4 | 5 | 3 | 1295/1771 | 9576/10626 | 25901 | 1 | $q$'s $= 0$ |

Table 4.2: Results for Brier's $\hat{C}$

from cluster sampling. Rao and Scott [77] suggest three method of moments estimators of $C$ for the more general case where $r_i$ items are observed at the $i^{\text{th}}$ group, $i = 1, \ldots, N$. The first of these simplifies to Brier's $\hat{C}$ when $r_i = r \ \forall i$. Wilson [91] derived a generalised least-squares estimator of $C$ that, again, simplifies to Brier's $\hat{C}$ and together Koehler and Wilson [55] suggested two further generalised least-squares estimators. For the case $r = 2$, Cohen [18] suggested estimating the quantity $C-1$ by maximum likelihood, starting from method of moments estimates. However, there it is assumed that the data contains information on the exact classification of each of the individual $2N$ items. Janardan and Patil [51] suggest another consistent method of moments estimator of $C$ based on data from only $K$ of the $K+1$ categories, but one which unattractively does vary depending on this choice, unlike Mosimann's $\hat{C}$. For the Beta-Binomial distribution ($K = 1$), Kleinman [54] produced (4.7) and estimated the quantity $1/(\tau+1)$ by method of moments, as opposed to estimating $C$. Danaher [21] accepted (4.7) for estimating the parameters of the Dirichlet-Multinomial distribution and suggested a way of estimating $\tau$ in the case $r = 1$ by assuming the availability of supplementary Beta-Binomial data pertaining to the marginals of the Dirichlet-Multinomial distribution.

## 4.2 Pseudo Maximum Likelihood Estimates

This method may be useful when the probability distribution of the data depends on two mutually exclusive sets of parameters $\eta_1$ and $\eta_2$ and the likelihood surface is ill-behaved for the full problem, but well-behaved for the restricted problem. If an estimate $\hat{\eta}_1$ of $\eta_1$ (*not* the maximum likelihood estimate) can be obtained, then $\hat{\eta}_2$, the solution of the 'pseudo likelihood' equations

$$\frac{\partial}{\partial \eta_2} L(\hat{\eta}_1, \eta_2 \,|\, \text{data}) = 0\,,$$

is called the pseudo maximum likelihood estimate of $\eta_2$. Gong and Samaniego [37] give conditions under which $\hat{\eta}_2$ is consistent and asymptotically normal.

Pseudo maximum likelihood estimation of the parameter $\underline{\alpha}$ in the Dirichlet-Multinomial distribution has been found by Chuang and Cox [17] to constitute a good compromise between full maximum likelihood (discussed in §4.5.2) and method of moments. As has already been noted, estimation of $\underline{\alpha}$ falls naturally into two

parts: estimation of the mean parameters $\lambda_j \equiv \alpha_j/\tau$, $j = 1, \ldots, K$, and the scale parameter $\tau$. Chuang and Cox showed that the standard deviation of the moment estimator (4.7), or

$$\hat{\lambda}_j = \frac{\hat{\alpha}_j}{\hat{\tau}} = \frac{\bar{Y}_j}{r} \tag{4.20}$$

where $1 \leq j \leq K$, is bounded above by 0.5 and suggested that (4.20) is an adequate estimate of $\lambda_j$. Pseudo maximum likelihood estimation of $\tau$ is appealing due to a conjecture of Good [38], which was proved in a more general context by Levin and Reeds [68] to give the following result.

**Theorem 4.2.1** *Let $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$ be observations from DMD$(r, \underline{\alpha})$, where $\underline{\alpha} = \tau\underline{\lambda}$ $\left(\tau = \sum_{j=1}^{K+1} \alpha_j\right)$. Assume $\underline{\lambda} > \underline{0}$ is known. Then*

$$L(\tau) \equiv \prod_{i=1}^{N} L\left(\tau\underline{\lambda} \middle| \underline{Y}^{(i)}\right)$$

*has at most one local maximum. It occurs at finite $\tau$ if*

$$X^2 > \sum_{j=1}^{K+1} \frac{N\bar{Y}_j}{\lambda_j r} - N$$

*and at $\tau = \infty$ otherwise, where*

$$X^2 = \sum_{i=1}^{N} \sum_{j=1}^{K+1} \frac{\left(Y_j^{(i)} - r\lambda_j\right)^2}{r\lambda_j}.$$

PROOF: See [68]. □

Replacing $\underline{\lambda}$ in the above result with $\hat{\underline{\lambda}}$ given by (4.20) it follows that a unique finite pseudo maximum likelihood estimate of $\tau$ exists whenever

$$X^2 = \sum_{i=1}^{N} \sum_{j=1}^{K+1} \frac{\left(Y_j^{(i)} - \bar{Y}_j\right)^2}{\bar{Y}_j} > NK\,. \tag{4.21}$$

In this case, $X^2$ is Pearson's $\chi^2$-statistic for testing equality of $\underline{\theta}^{(i)}$, $i = 1, \ldots, N$. Note also that if (4.21) is not satisfied for a given set of data, the resultant pseudo maximum likelihood estimate, $\tau = \infty$, suggests that the Multinomial distribution may have been more appropriate for modelling the data than the Dirichlet-Multinomial. Paul *et al.* [73] derive and discuss tests for goodness of fit of the Multinomial distribution against Dirichlet-Multinomial alternatives.

**Example 4.2.1** *To illustrate the calculation of $\hat{\tau}$ by pseudo maximum likelihood, consider again the data from Mosimann [72], reproduced in Table B.1 of Appendix B, §B.1. For this data, recall $N = 73$, $r = 100$, $K + 1 = 4$ and (4.14), from Example 4.1.1. Then, by (4.20),*

$$\begin{aligned}
\hat{\lambda}_1 &= \frac{6298/73}{100} = 0.863 \quad (3 \text{ d.p.}) \\
\hat{\lambda}_2 &= \frac{103/73}{100} = 0.014 \quad (3 \text{ d.p.}) \\
\hat{\lambda}_3 &= \frac{662/73}{100} = 0.091 \quad (3 \text{ d.p.}) \\
\hat{\lambda}_4 &= \frac{237/73}{100} = 0.032 \quad (3 \text{ d.p.}).
\end{aligned}$$

*Solving*

$$\frac{\partial}{\partial \tau}\left[\prod_{i=1}^{N} L\left(\tau\hat{\underline{\lambda}}\middle|\underline{Y}^{(i)}\right)\right] = 0$$

*amounts to solving*

$$\sum_{i=1}^{N}\sum_{j=1}^{K+1}\sum_{l=0}^{Y_j^{(i)}-1}\frac{\hat{\lambda}_j}{l+\tau\hat{\lambda}_j} = N\sum_{l=0}^{r-1}\frac{1}{l+\tau}$$

*(see Danaher [21]) and gives $\hat{\tau} = 62.969$ (3 d.p.).*

All of the previously mentioned methods for estimating $\underline{\alpha}$ employ (4.7), but differ in their procedure for estimating $C$ or, equivalently, $\tau$ directly. However, there is no need to repeat the type of coherency analysis carried out for Mosimann's and Brier's $\hat{C}$ as the following development will demonstrate.

## 4.3 Solving for $\hat{\tau}$

The widespread acceptance of (4.20) in estimating $\underline{\alpha}$ and the desire to learn which accompanying estimates of $\tau$ lead to coherent predictive inference suggest the following approach. For a given set of conditioning histograms, $\mathcal{H}$, derive the system of coherency induced equations using (4.3), where $\hat{\underline{\alpha}} = \hat{\tau}\hat{\underline{\lambda}}$ with $\hat{\underline{\lambda}}$ given by (4.20), but $\hat{\tau}$ remains an unknown variable. Of course, there will actually be many different $\hat{\tau}$ variables corresponding to estimates of $\underline{\alpha}$ from different histograms. Then solve this system of homogeneous equations for the $\hat{\tau}$ values and the components of $\underline{q}$ appearing in the equations, subject to the constraint that all of these components of

$\underline{q}$ are positive. Such a procedure would find all of the possible ways of estimating $\tau$ that combine with (4.20) to produce predictive probabilities that may be coherently assessed with a nonzero probability of being used.

Let $\mathcal{H}_P$ be the set consisting of all positive histograms. Clearly, $\mathcal{H}_P$ is empty unless $rN \geq K+1$. It seems sensible to adopt $\mathcal{H}_P$ as the set of conditioning histograms in the above approach as this is the largest such set that will produce valid estimates using (4.20). Recognise though that $\mathcal{H}_P$ is generally larger than either $\mathcal{H}_M$ or $\mathcal{H}_B$. Before proceeding, we develop a number of results concerning $\mathcal{H}_P$.

**Lemma 4.3.1** *The cardinality of $\mathcal{H}_P$,* i.e., *the number of positive histograms, is*

$$ {}^{(N+{}^{r+K}C_K-1)}C_N - \sum_{j=1}^{K}\left[(-1)^{j+1}\left({}^{K+1}C_j\right)\left({}^{(N+{}^{r+K-j}C_{K-j}-1)}C_N\right)\right]. $$

PROOF: Fix $J \in \{1,\ldots,K\}$. For any $J$ categories, the number of types with no items in any of these categories is the number of ways of putting $r$ things into $K+1-J$ boxes, namely ${}^{r+K-J}C_{K-J}$. Therefore the number of histograms with no items in any of these $J$ categories is the number of ways of putting $N$ things into ${}^{r+K-J}C_{K-J}$ boxes, namely ${}^{(N+{}^{r+K-J}C_{K-J})}C_N$. Hence, by the principal of inclusion and exclusion (see [29]), the number of histograms with at least one item in each category is

$$ {}^{(N+{}^{r+K}C_K-1)}C_N - \sum_{j=1}^{K}\left[(-1)^{j+1}\left({}^{K+1}C_j\right)\left({}^{(N+{}^{r+K-j}C_{K-j}-1)}C_N\right)\right]. $$

□

**Lemma 4.3.2** *For fixed $r$ and $K+1$, the proportion of positive histograms out of the total number of possible histograms tends to 1 as $N$ increases.*

PROOF: By Lemma 4.3.1, the proportion of positive histograms is

$$ 1 - \left(\sum_{j=1}^{K}\left[(-1)^{j+1}\left({}^{K+1}C_j\right)\left({}^{(N+{}^{r+K-j}C_{K-j}-1)}C_N\right)\right]\right) \Big/ {}^{(N+{}^{r+K}C_K-1)}C_N, $$

which, for large $N$, behaves like

$$ 1 - \left(\sum_{j=1}^{K}\left[(-1)^{j+1}\left({}^{K+1}C_j\right)\frac{N^{({}^{r+K-j}C_{K-j}-1)}}{({}^{r+K-j}C_{K-j}-1)!}\right]\right) \Big/ \left(\frac{N^{({}^{r+K}C_K-1)}}{({}^{r+K}C_K-1)!}\right) $$

$$ \sim 1, \quad \text{for large} N. $$

**Lemma 4.3.3** *For $N > (K+1)/r$, the set $\mathcal{H}_P$ of all positive histograms is linked.*

PROOF: Case (a) $r \geq K+1$
Take any two distinct histograms, $h$ and $\mathfrak{h} \in \mathcal{H}_P$. Let $y_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $\mathfrak{h}$ and suppose type $m$ is

$$\begin{pmatrix} r-K \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Consider the following algorithm.

i. Let $Hu = h$.

ii. If $d(Hu, \mathfrak{h}) = 0$, stop. Otherwise, let $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denote the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $Hu$. If $x_m = 0$, go to Step iii. If $x_m > 0$, since

$$\sum_{t=1}^{{}^{r+K}C_K} x_t = \sum_{t=1}^{{}^{r+K}C_K} y_t = N\,,$$

there exist $s_1, s_2 \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $x_{s_1} > y_{s_1}$, $x_{s_2} < y_{s_2}$. If possible, choose $s_1 \neq m$. Link $Hu$ to $Hv$, where $Hv$ is defined to be the histogram that has one less type $s_1$ and one more type $s_2$ than $Hu$ does. These histograms may be written

$$\begin{aligned} x_1, \ldots, x_{s_1}, \ldots, x_{s_2}, \ldots, x_{r+K C_K} \qquad & Hu \\ x_1, \ldots, x_{s_1} - 1, \ldots, x_{s_2} + 1, \ldots, x_{r+K C_K} \qquad & Hv \\ y_1, \ldots, y_{s_1}, \ldots, y_{s_2}, \ldots, y_{r+K C_K} \qquad & \mathfrak{h}\,, \end{aligned}$$

assuming, without loss of generality, that $s_1 < s_2$. Obviously, $d(Hv, \mathfrak{h}) = d(Hu, \mathfrak{h}) - 2$. Go to Step iv.

iii. Since $Hu \neq h$ and

$$\sum_{t=1}^{{}^{r+K}C_K} x_t = \sum_{t=1}^{{}^{r+K}C_K} y_t = N\,,$$

there exists $s_1 \in \left\{1, \ldots, {}^{r+K}C_K\right\}$ such that $x_{s_1} > y_{s_1}$. Link $Hu$ to $Hv$, where $Hv$ is defined to be the histogram that has one less type $s_1$ than $Hu$ does and

one type $m$. These histograms may be written

$$\begin{aligned} x_1, \ldots, x_{s_1}, \ldots, x_m = 0, \ldots, x_{r+KC_K} \qquad & Hu \\ x_1, \ldots, x_{s_1} - 1, \ldots, 1, \ldots, x_{r+KC_K} \qquad & Hv \\ y_1, \ldots, y_{s_1}, \ldots, y_m, \ldots, y_{r+KC_K} \qquad & \mathfrak{h}\,, \end{aligned}$$

assuming, without loss of generality, that $s_1 < m$. Either, $d(Hv, \mathfrak{h}) = d(Hu, \mathfrak{h})$ or $d(Hv, \mathfrak{h}) = d(Hu, \mathfrak{h}) - 2$.

iv. Let $Hu = Hv$. Go to Step ii.

The algorithm is guaranteed to terminate because the distance of the current histogram from $\mathfrak{h}$ is being reduced each time through, except possibly once if Step ii leads to Step iii, and is bounded below by 0. Since $h$ and $\mathfrak{h}$ are distinct, $d(h, \mathfrak{h}) \neq 0$ so that the algorithm will not stop the first time Step ii is encountered. If it stops the second time Step ii is encountered, then $h$ and $\mathfrak{h}$ are directly linked. Otherwise, $h$ and $\mathfrak{h}$ are linked indirectly through the sequence of histograms $\{Hv\}$. The only point that needs to be verified is that all of these 'intermediate' histograms are also members of $\mathcal{H}_P$, *i.e.*, positive histograms. This is evident because any intermediate histograms contain a type $m$, ensuring that at least one item from the $N$ groups has been observed in each category.

Case (b) $r < K + 1$

For the purpose of this proof, consider a histogram to be identified by the $(K+1) \times N$ matrix whose columns represent the type outcomes of the $N$ groups, these being ordered so that all those of type 1 come first, followed by all those of type 2, and so on. In this light, two histograms are linked if and only if exactly $N - 1$ of their columns are the same, and the distance between any two histograms, $h$ and $\mathfrak{h}$, is defined to be

$$\mathfrak{d}(h, \mathfrak{h}) = \sum_{i=1}^{N} \sum_{j=1}^{K+1} (h_{ji} \neq \mathfrak{h}_{ji})\,,$$

*i.e.*, the number of matrix entries in which $h$ and $\mathfrak{h}$ differ. If a histogram, $h \in \mathcal{H}_P$, has $h_{JI} > 1$ for some $J \in \{1, \ldots, K+1\}, I \in \{1, \ldots, N\}$, it can be linked to a histogram $h' \in \mathcal{H}_P$ whose $I^{\text{th}}$ column has $h'_{JI} = 1$ and for $(h_{JI} - 1)$ of the indices from the set $\{j : h_{jI} = 0, j \in \{1, \ldots, K+1\}\}$, $h'_{jI} = 1$. Thus we need only consider the following. Take any two distinct histograms, $h$ and $\mathfrak{h} \in \mathcal{H}_P$, with $h_{ji}, \mathfrak{h}_{ji} \in \{0, 1\}\ \forall\ j \in \{1, \ldots, K+1\}, i \in \{1, \ldots, N\}$. For a given histogram,

define a *floating* 1 to be a 1 that occurs in a row containing at least two 1's. Since $N > (K+1)/r$, every positive histogram contains a floating 1. Put a box around all the entries, $h_{ji}$, in $h$ for which $h_{ji} = \mathfrak{h}_{ji}$. Note that for any column $i, i \in \{1, \ldots, N\}$, of $h$,

$$\underbrace{\sum_{j=1}^{K+1} h_{ji}}_{\boxed{h_{ji}} \text{ boxed}} + \underbrace{\sum_{j=1}^{K+1} h_{ji}}_{h_{ji} \text{ unboxed}} = r$$

and

$$\underbrace{\sum_{j=1}^{K+1} h_{ji}}_{\boxed{h_{ji}} \text{ boxed}} + \underbrace{\sum_{j=1}^{K+1} [1 - h_{ji}]}_{h_{ji} \text{ unboxed}} = r$$

(the latter equation being the summation of the $i^{\text{th}}$ column of $\mathfrak{h}$). Hence

$$\underbrace{\sum_{j=1}^{K+1} h_{ji}}_{h_{ji} \text{ unboxed}} = \underbrace{\sum_{j=1}^{K+1} [1 - h_{ji}]}_{h_{ji} \text{ unboxed}},$$

which implies that each column of $h$ contains the same number of unboxed 1's as it does unboxed 0's. Consider the following algorithm.

i. Let $Hu = h$.

ii. If $\mathfrak{d}(Hu, \mathfrak{h}) = 0$, stop. If there are no unboxed floating 1's in $Hu$, go to Step iii. Otherwise, convert an unboxed floating 1 in $Hu$ into a boxed $\boxed{0}$ and convert an unboxed 0 in the same column into a boxed $\boxed{1}$. This creates a new histogram, $Hv$, linked to $Hu$, with $\mathfrak{d}(Hv, \mathfrak{h}) = \mathfrak{d}(Hu, \mathfrak{h}) - 2$. Go to Step v.

iii. If there are no unboxed 1's in $Hu$ in the same column as a boxed floating $\boxed{1}$, go to Step iv. Otherwise, convert a boxed floating $\boxed{1}$, that is in the same column as an unboxed 1, into an unboxed 0 and convert an unboxed 0 in the same column into a boxed $\boxed{1}$. This creates a new histogram, $Hv$, linked to $Hu$, with $\mathfrak{d}(Hv, \mathfrak{h}) = \mathfrak{d}(Hu, \mathfrak{h})$. However, the row in which the boxed $\boxed{1}$ was just created already contained an unboxed 1, which is now an unboxed floating 1. Go to Step v.

iv. Convert a boxed floating $\boxed{1}$ in $Hu$ into an unboxed 0 and convert a boxed $\boxed{0}$ in the same column, that is in the same row as an unboxed $1^*$, into an unboxed 1. This creates a new histogram, $Hv$, linked to $Hu$, with $\mathfrak{d}(Hv, \mathfrak{h}) = \mathfrak{d}(Hu, \mathfrak{h}) + 2$.

However, the original unboxed 1* in the row in which the latest unboxed 1 was just created is now an unboxed floating 1.

v. Let $Hu = Hv$. Go to Step ii.

The algorithm is guaranteed to terminate because the distance of the current histogram from $\mathfrak{h}$ is being reduced each time through, except possibly if Step ii leads to Step iii, and is bounded below by 0. However, if Step iii *is* encountered and does not lead to Step iv, the next move is guaranteed to yield, via Step v, a successful completion of Step ii. If Step iii is encountered and does lead to Step iv, the next move is guaranteed (with the right choice of floating 1* suggested in Step iv) to yield, via Step v, a successful completion of Step ii and then again, via Step v, a successful completion of Step ii. Since $h$ and $\mathfrak{h}$ are distinct, $d(h, \mathfrak{h}) \neq 0$ so that the algorithm will not stop the first time Step ii is encountered. If it stops the second time Step ii is encountered, then $h$ and $\mathfrak{h}$ are directly linked. Otherwise, $h$ and $\mathfrak{h}$ are linked indirectly through the sequence of histograms $\{Hv\}$. It needs to be verified that all of these 'intermediate' histograms are also members of $\mathcal{H}_P$, *i.e.*, positive histograms. This is evident because a column containing a floating 1, in the $j^{\text{th}}$ category, say, can obviously be replaced by a column that has a 0 in its $j^{\text{th}}$ category, without affecting positivity. For completeness, the existence of the unboxed 1 referred to in the penultimate statement of Step iii and the boxed $\boxed{0}$ referred to in the first statement of Step iv also needs to be proven. Suppose there are unboxed 1's in $Hu$, but no unboxed floating 1's. Then in any row containing an unboxed 1, all of the other entries are 0's and at least one of these is an unboxed 0 since $\mathfrak{h} \in \mathcal{H}_P$. Suppose $Hu$ contains $z$ unboxed 1's, and therefore $z$ unboxed 0's, in total. Then within $Hu$ there are $z$ rows each containing just one unboxed 1. But each of these $z$ rows contains at least one unboxed 0, and therefore each contains exactly one unboxed 0. (Note that $z \geq 2$.) Hence, there are only two kinds of rows in $Hu$. The first kind contain one unboxed 1, one unboxed 0 and $(N-2)$ boxed $\boxed{0}$'s, while the second kind contain $N$ boxed numbers including at least one boxed $\boxed{1}$. Suppose further that $Hu$ has no unboxed 1's in the same column as a boxed floating $\boxed{1}$. Then there are no unboxed 0's in the same column as a boxed floating $\boxed{1}$ either. Hence there is a boxed $\boxed{0}$ in the same column as a boxed floating $\boxed{1}$ in a row of the first kind.

Note that, in practice, case (a) of the above proof is the more likely situation as it applies whenever the number of items sampled from each group is at least as

great as the number of categories into which they may be classified. □

We now return to the issue of leaving the values of $\hat{\tau}$ unspecified in the conditioning equations and solving for them. One method would be to assign a different $\hat{\tau}$ variable to each of the possible positive histograms. However, this would be unnecessarily complicated. The value $\hat{\tau}$ can be thought of as measuring strength of belief in the corresponding estimate $\underline{\hat{\alpha}}$ and, as such, pertains to overall information provided by the observations without regard to the *ordering* of the categories into which they were classified. Hence, outcomes of $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$ that differ only due to the ordering of the categories constitute histograms that should be assigned the same $\hat{\tau}$ variable. This may be most easily determined by again considering a histogram to be identified by the $(K+1) \times N$ matrix whose columns represent the type outcomes of the $N$ groups. Then, two histograms should be assigned the same $\hat{\tau}$ variable if one of them can be obtained from the other by permuting rows (and columns, if necessary).

**Example 4.3.1** *Let $N+1=3$, $r=2$, $K+1=3$. The histograms formed by*

$$\left\{\underline{Y}^{(1)}, \underline{Y}^{(2)}, \underline{Y}^{(3)}\right\} = \left\{\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right\}$$

*and*

$$\left\{\underline{Y}^{(1)}, \underline{Y}^{(2)}, \underline{Y}^{(3)}\right\} = \left\{\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right\}$$

*can be considered to be*

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 2 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

*respectively, and should be assigned the same $\hat{\tau}$ variable. The two histograms contain* different *information, but the same information* strength.

Use of the approach outlined at the beginning of §4.3 was accomplished by developing an algorithm which takes the values $N+1$, $r$ and $K+1$ as input, finds the set of histograms, $\mathcal{H}_P$, and derives the system of coherency induced equations

involving the $\hat{\tau}$'s as extra variables. The algorithm was implemented in MAPLE V RELEASE 3 (see Appendix A, §A.4) and the SOLVE routine of this package was used to find those valid $\hat{\tau}$ values that allow a nonzero solution for the components of $\underline{q}$ that appear in the resulting system of equations. The results for problems of various sizes are presented in Table 4.3, with an accompanying key on pages 93 and 94. Due to the added complexity of solving equations which are nonlinear in the $\hat{\tau}$ variables, again only small-sized problems were practical for consideration by this method. The following explanations apply to expressions used in Table 4.3.

- taucount — Number of distinct $\hat{\tau}$ variables.
- histcount — Number of histograms in $\mathcal{H}_P$.
- dimhist — Number of histograms in $\mathcal{H}'$, *i.e.*, total number of possible histograms.
- qcount — Number of variables, *i.e.*, components of $\underline{q}$, involved in the system of equations generated by histograms from $\mathcal{H}_P$.
- dimq — Total number of variables, *i.e.*, components of $\underline{q}$.
- eqncount — Number of coherency induced equations, including the condition that all of the variables sum to 1 ($\underline{q}^T\underline{1} = 1$).
- $\hat{\tau}$'s = # — Unless otherwise specified, all of the $\hat{\tau}$ variables must equal the value # to allow a nonzero solution for the components of $\underline{q}$ that appear in the system of equations generated by histograms from $\mathcal{H}_P$.
- * — As opposed to solving the full system of coherency induced equations, these results were found by solving for $\hat{\tau}$ values that satisfy all 3-cycles generated by histograms from $\mathcal{H}_P$ (recall Theorem 2.4.1 and see Theorem 4.3.4 and Appendix A, §A.5). It is important to note that the results were obtained very quickly using this approach, whereas solving the full system of equations would have been extremely time consuming, if not completely impractical.

The results in Table 4.3 suggest that if you want to be coherent in giving a nonzero probability to using a strategy that estimates your predictive probabilities

| $N+1$ | $r$ | $K+1$ | taucount | histcount/ dimhist | qcount/ dimq | eqncount | Solution |
|---|---|---|---|---|---|---|---|
| 3 | 2 | 3 | 2 | 6/21 | 28/56 | 31 | $\hat{\tau}_1 = \hat{\tau}_1, \hat{\tau}_2 = \hat{\tau}_2$ |
| 4 | 2 | 3 | 8 | 29/56 | 84/126 | 146 | $\hat{\tau}_1 = \frac{18\hat{\tau}_4}{\hat{\tau}_4+12}, \hat{\tau}_4 = \hat{\tau}_4, \hat{\tau}_5 = \hat{\tau}_5, \hat{\tau}$'s $= 6$ |
| 5 | 2 | 3 | 19 | 84/126 | 192/252 | 421 | $\hat{\tau}_6 = \hat{\tau}_6, \hat{\tau}$'s $= 8$ |
| 6 | 2 | 3 | 40 | 192/252 | 381/462 | 961 | $\hat{\tau}_6 = \frac{25\hat{\tau}_{18}}{\hat{\tau}_{18}+15}, \hat{\tau}_{18} = \hat{\tau}_{18}, \hat{\tau}$'s $= 10$ |
| 7 | 2 | 3 | 77 | 381/462 | 687/792 | 1906 | $\hat{\tau}_6 = \frac{210\hat{\tau}_{46}}{13\hat{\tau}_{46}+54}, \hat{\tau}_{19} = \frac{30\hat{\tau}_{46}}{\hat{\tau}_{46}+18}, \hat{\tau}_{46} = \hat{\tau}_{46}, \hat{\tau}$'s $= 12$ |
| 3 | 3 | 3 | 7 | 28/55 | 162/220 | 253 | $\hat{\tau}$'s $= 6$ |
| 4 | 3 | 3 | 34 | 163/220 | 613/715 | 1468 | $\hat{\tau}$'s $= 9$ |
| 3 | 3 | 4 | 5 | 46/210 | 684/1540 | 875 | $\hat{\tau}_4 = \hat{\tau}_4, \hat{\tau}$'s $= 6$ * |
| 3 | 4 | 3 | 17 | 78/120 | 577/680 | 1093 | $\hat{\tau}$'s $= 8$ * |
| 3 | 4 | 4 | 19 | 236/630 | 5022/7770 | 8025 | $\hat{\tau}$'s $= 8$ * |

Table 4.3: Results of Solving for $\hat{\tau}$

as in (4.3) where

$$\hat{\alpha}_j = \frac{\bar{Y}_j}{r}\hat{\tau}, \qquad j = 1, \ldots, K+1, \tag{4.22}$$

then you are, in fact, very restricted in the way you estimate $\tau$. Specifically, for given values of $N+1$, $r$ and $K+1$ it seems that your estimates of $\tau$ corresponding to most sorts of positive histograms must all equal $rN$. Furthermore, if $r \geq K+1$ this may be true for *all* positive histograms.

**Key for Table 4.3**:

For given $N+1$, $r$ and $K+1$ it would be tedious to detail what sort of histogram corresponds to each and every one of the $\hat{\tau}$ variables. However, it may be of interest to know what sort of histograms correspond to those $\hat{\tau}$ variables that are not completely determined in the 'Solution' column, these generally being exceptional cases. Table 4.4 contains this information.

For the examples in Table 4.3 where $r < K+1$, the sorts of histograms corresponding to $\hat{\tau}$ variables that are not forced to equal $rN$ appear to represent outcomes of the $N$ groups for which $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$ are quite dissimilar.

Note that (4.22) and $\hat{\tau} = rN$ together imply

$$\begin{aligned} \hat{\alpha}_j &= \frac{\bar{Y}_j}{r}\hat{\tau} \\ &= N\bar{Y}_j \\ &= \sum_{i=1}^{N} Y_j^{(i)}, \qquad j = 1, \ldots, K+1, \end{aligned} \tag{4.23}$$

which is the total number of observed items falling in the $j^{\text{th}}$ category. Recalling that $\hat{\tau}$ may be interpreted as the strength of belief in the estimate $\underline{\hat{\alpha}}$, $\hat{\tau} = rN$ implies that (4.23) is, in some sense, given the same weighting as the amount of data already seen, in contributing to your predictive probabilities.

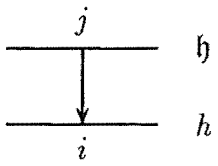
**Theorem 4.3.4** *If you assert your predictive probabilities using (4.3) where* $\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)}$, $j = 1, \ldots, K+1$, *for all* $H \in \mathcal{H}_P$, *the set of positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Consider a given $n$-cycle formed by histograms from $\mathcal{H}_P$. Every histogram involved in the $n$-cycle appears as a subscript exactly once in the numerator and

| $N+1$ | $r$ | $K+1$ | $\hat{\tau}$ Assignment |
|---|---|---|---|
| 3 | 2 | 3 | $\tau_1 \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \tau_2 \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$ |
| 4 | 2 | 3 | $\tau_1 \begin{bmatrix} 2 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \tau_4 \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \tau_5 \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ |
| 5 | 2 | 3 | $\tau_6 \begin{bmatrix} 2 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$ |
| 6 | 2 | 3 | $\tau_6 \begin{bmatrix} 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad \tau_{18} \begin{bmatrix} 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$ |
| 7 | 2 | 3 | $\tau_6 \begin{bmatrix} 2 & 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad \tau_{19} \begin{bmatrix} 2 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$ $\tau_{46} \begin{bmatrix} 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 \end{bmatrix}$ |
| 3 | 3 | 4 | $\tau_4 \begin{bmatrix} 2 & 0 \\ 1 & 0 \\ 0 & 2 \\ 0 & 1 \end{bmatrix}$ |

Table 4.4: Key for Some $\hat{\tau}$ Variables in Table 4.3

once in the denominator of the overall $n$-cycle probability ratio. Similarly, every type involved in the $n$-cycle appears an equal number of times as a subscript in this numerator and denominator. Hence, only the terms in (4.3) that depend simultaneously on the histogram *and* type will determine whether or not this $n$-cycle is satisfied. As we have already discovered, each occurrence




in the $n$-cycle contributes $p_{i,h}/p_{j,\mathfrak{h}}$ to the relevant ratio of probabilities, due to the fact that histogram $\mathfrak{h}$ contains one less type $j$ and one more type $i$ than does $h$. Suppose types $i$ and $j$ are

$$\begin{pmatrix} a_1 \\ \vdots \\ a_{K+1} \end{pmatrix} \text{ and } \begin{pmatrix} b_1 \\ \vdots \\ b_{K+1} \end{pmatrix},$$

respectively. For $j = 1, \ldots, K+1$, let $S_{j,h}$ and $S_{j,\mathfrak{h}}$ be $\sum_{i=1}^{N} Y_j^{(i)}$ of histogram $h$ and $\mathfrak{h}$, respectively. Then

$$p_{i,h} = \frac{r!\Gamma(rN)\prod_{j=1}^{K+1}\Gamma(S_{j,h}+a_j)}{\prod_{j=1}^{K+1}[a_j!]\prod_{j=1}^{K+1}[\Gamma(S_{j,h})]\Gamma(rN+r)}$$

and

$$p_{j,\mathfrak{h}} = \frac{r!\Gamma(rN)\prod_{j=1}^{K+1}\Gamma(S_{j,\mathfrak{h}}+b_j)}{\prod_{j=1}^{K+1}[b_j!]\prod_{j=1}^{K+1}[\Gamma(S_{j,\mathfrak{h}})]\Gamma(rN+r)}.$$

However, for the reasons given above, we need only look at the ratio of those terms in $p_{i,h}$ and $p_{j,\mathfrak{h}}$ that depend simultaneously on histogram and type, namely

$$\begin{aligned} \frac{\prod_{j=1}^{K+1}\Gamma(S_{j,h}+a_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,\mathfrak{h}}+b_j)} &= \frac{\prod_{j=1}^{K+1}\Gamma(S_{j,h}+a_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,h}-b_j+a_j+b_j)} \\ &= \frac{\prod_{j=1}^{K+1}\Gamma(S_{j,h}+a_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,h}+a_j)} \\ &= 1. \end{aligned}$$

Hence the overall ratio of predictive probabilities equals 1 and the $n$-cycle is satisfied. Lemma 4.3.3 and the fact that your probabilities are strictly positive then allows Theorem 2.4.1 to complete the proof. □

The following theorems verify the pattern of results displayed in Table 4.3 and prove the uniqueness of the solution $\hat{\tau} = rN$, given (4.22), when $r \geq K+1$. This is accomplished by finding 3-cycles or combinations of 3-cycles that are only satisfied if some of the $\hat{\tau}$ variables involved equal $rN$, and showing that *all* of the $\hat{\tau}$ variables corresponding to different sorts of histograms can be determined in this way. Satisfying these 3-cycles amounts to solving expressions of the form

$$\prod_{h=1}^{3} \left[ \frac{\prod_{j=1}^{K+1} \Gamma\left(\frac{z_j \hat{\tau}_h}{rN} + a_j\right)}{\prod_{j=1}^{K+1} \Gamma\left(\frac{z_j \hat{\tau}_h}{rN} + b_j\right)} \right] = 1\,, \tag{4.24}$$

where $z_j + a_j > 0, z_j + b_j > 0, j = 1, \ldots, K+1$, $z_{K+1} = rN - \sum_{j=1}^{K} z_j$ and $\sum_{j=1}^{K+1} a_j = \sum_{j=1}^{K+1} b_j = r$ with $a_J \neq b_J$ for some $J \in \{1, \ldots, K+1\}$. Using the recursive definition $\Gamma(x+1) = x\Gamma(x)$ to remove the Gamma expressions, the left-hand side of (4.24) simplifies to give a polynomial in some subset, $T$, of $\{\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3\}$ in the numerator and in the denominator. The roots of the difference in these two polynomials will always include the case where all $\hat{\tau} \in T$ equal $rN$. Generally, this is the only valid solution, *i.e.*, the only solution giving positive values to all $\hat{\tau} \in T$, though not always. It suffices to show, however, that all $\hat{\tau}$ variables do occur in a 3-cycle with this property. The following theorem due to Descartes will prove most useful in this respect.

**Theorem 4.3.5** *The number of positive real roots of a real polynomial is equal to the number of sign changes in its coefficients, or differs from it by a positive even number.*

PROOF: See [88]. □

We also make the following definition.

**Definition 4.3.1** *A* sub-histogram *is a histogram of types generated from the outcomes of $N-1$ groups.*

**Theorem 4.3.6** *Let $N+1 \geq 3, r \geq K+1$ and $K+1 \geq 3$. If you assert your predictive probabilities using (4.3) where*

$$\hat{\alpha}_j = \frac{\bar{Y}_j}{r}\hat{\tau}, \qquad j = 1, \ldots, K+1\,,$$

*for all $H \in \mathcal{H}_P$, the set of positive histograms, then all of the $\hat{\tau}$ variables must equal $rN$ to allow a nonzero solution for the components of $\underline{q}$ that appear in the system of coherency induced equations.*

PROOF: The complex nature of this proof will make it convenient to temporarily ignore the ordering that has previously been assigned to both types and histograms. It will also be easiest, at times, to return to the notion of a histogram being identified by the $(K+1) \times N$ matrix whose columns represent the type outcomes of the $N$ groups.

Consider the 3-cycle formed by the histograms $H1$, $H2$, $H3$ in $\mathcal{H}_P$ which are defined to be a given sub-histogram plus a type $k$, $l$ and $m$, respectively, these types being

$$\begin{pmatrix} a_1 \\ \vdots \\ a_{K+1} \end{pmatrix}, \begin{pmatrix} b_1 \\ \vdots \\ b_{K+1} \end{pmatrix} \text{and} \begin{pmatrix} c_1 \\ \vdots \\ c_{K+1} \end{pmatrix},$$

respectively, where $a_{K+1} = r - \sum_{j=1}^{K} a_j$, $b_{K+1} = r - \sum_{j=1}^{K} b_j$, $c_{K+1} = r - \sum_{j=1}^{K} c_j$. This 3-cycle may be represented as in Figure 4.1 and will be satisfied if and only if

$$\frac{p_{l,H1}p_{m,H2}p_{k,H3}}{p_{m,H1}p_{k,H2}p_{l,H3}} = 1\,. \tag{4.25}$$

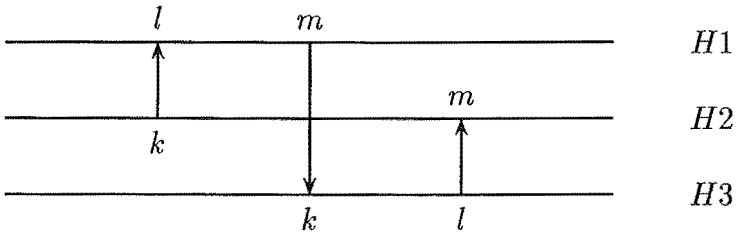



Figure 4.1: Line Representation of a 3-Cycle

Suppose that $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_3$ represent the $\hat{\tau}$ variables associated with $H1$, $H2$ and $H3$, respectively, each scaled by a factor of $1/rN$. Hence, we wish to prove, eventually, that to satisfy all 3-cycles requires $\hat{\tau}_1 = \hat{\tau}_2 = \hat{\tau}_3 = 1$. Suppose also that the sub-histogram has

$$\sum_{i=1}^{N-1} Y_j^{(i)} = z_j, \qquad j = 1, \ldots, K+1\,,$$

where $z_j + a_j, z_j + b_j, z_j + c_j > 0$, $j = 1, \ldots, K+1$ and $z_{K+1} = r(N-1) - \sum_{j=1}^{K} z_j$. Remembering that only the terms in the predictive probabilities of an $n$-cycle ratio that depend simultaneously on histogram and type will determine whether or not

the $n$-cycle is satisfied, (4.25) may be rewritten

$$\frac{\prod_{j=1}^{K+1}\Gamma(S_{j,H1}\hat{\tau}_1+b_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,H1}\hat{\tau}_1+c_j)}\frac{\prod_{j=1}^{K+1}\Gamma(S_{j,H2}\hat{\tau}_2+c_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,H2}\hat{\tau}_2+a_j)}\frac{\prod_{j=1}^{K+1}\Gamma(S_{j,H3}\hat{\tau}_3+a_j)}{\prod_{j=1}^{K+1}\Gamma(S_{j,H3}\hat{\tau}_3+b_j)}=1\,, \qquad (4.26)$$

where for $j=1,\ldots,K+1$, $S_{j,H1}$, $S_{j,H2}$ and $S_{j,H3}$ are $\sum_{i=1}^{N}Y_j^{(i)}$ of histogram $H1$, $H2$ and $H3$, respectively. Suitable choices of the types $k$, $l$ and $m$ will now be made to produce various 3-cycles.

Firstly, let types $k$, $l$ and $m$ be

$$\begin{pmatrix}1\\0\\u_3\\u_4\\\vdots\\u_{K+1}\end{pmatrix},\begin{pmatrix}0\\1\\u_3\\u_4\\\vdots\\u_{K+1}\end{pmatrix}\text{ and }\begin{pmatrix}2\\0\\u_3-1\\u_4\\\vdots\\u_{K+1}\end{pmatrix},$$

respectively, where $u_3\in\{1,\ldots,r-1\}$ and $u_{K+1}=r-1-\sum_{j=3}^{K}u_j$ ($\sum_{j=3}^{K}u_j=0$ if $K=2$). Note we must assume that $z_1,z_2>0$ and, if $u_3=1$, that $z_3>0$. Then (4.26) simplifies to

$$\frac{z_2\hat{\tau}_1((z_3+u_3)\hat{\tau}_1+u_3-1)}{((z_1+1)\hat{\tau}_1+1)(z_1+1)\hat{\tau}_1}\frac{(z_1\hat{\tau}_2+1)}{((z_3+u_3)\hat{\tau}_2+u_3-1)}\frac{(z_1+2)\hat{\tau}_3}{z_2\hat{\tau}_3}=1$$

or

$$\frac{(z_1+2)((z_3+u_3)\hat{\tau}_1+u_3-1)(z_1\hat{\tau}_2+1)}{(z_1+1)((z_1+1)\hat{\tau}_1+1)((z_3+u_3)\hat{\tau}_2+u_3-1)}=1\,. \qquad (4.27)$$

If $(z_3+u_3)/z_1=u_3-1$, then Equation (4.27) becomes

$$\frac{(z_1+2)(z_1\hat{\tau}_1+1)}{(z_1+1)((z_1+1)\hat{\tau}_1+1)}=1\,,$$

with solution

$$\hat{\tau}_1=1\,.$$

If $(z_3+u_3)/(z_1+1)=u_3-1$, then Equation (4.27) becomes

$$\frac{(z_1+2)(z_1\hat{\tau}_2+1)}{(z_1+1)((z_1+1)\hat{\tau}_2+1)}=1\,,$$

with solution

$$\hat{\tau}_2=1\,.$$

Otherwise, Equation (4.27) gives

$$\hat{\tau}_1 = \frac{((z_1+1)(z_3+u_3) - z_1(z_1+2)(u_3-1))\hat{\tau}_2 - (u_3-1)}{-(z_3+u_3)\hat{\tau}_2 + (z_1+2)(z_3+u_3) - (z_1+1)^2(u_3-1)}. \tag{4.28}$$

Now let types $k$ and $l$ be as before, but type $m$ be

$$\begin{pmatrix} 0 \\ 2 \\ u_3 - 1 \\ u_4 \\ \vdots \\ u_{K+1} \end{pmatrix}.$$

Then $H1$, $H2$ and $\hat{\tau}_1$, $\hat{\tau}_2$ are the same histograms and variables as before, but $H3$ is different and $\hat{\tau}_3$ may or may not be. In this case, (4.26) simplifies to

$$\frac{((z_3+u_3)\hat{\tau}_1 + u_3 - 1)}{(z_2\hat{\tau}_1+1)} \frac{((z_2+1)\hat{\tau}_2+1)(z_2+1)\hat{\tau}_2}{z_1\hat{\tau}_2((z_3+u_3)\hat{\tau}_2+u_3-1)} \frac{z_1\hat{\tau}_3}{(z_2+2)\hat{\tau}_3} = 1$$

or

$$\frac{(z_2+1)((z_3+u_3)\hat{\tau}_1+u_3-1)((z_2+1)\hat{\tau}_2+1)}{(z_2+2)(z_2\hat{\tau}_1+1)((z_3+u_3)\hat{\tau}_2+u_3-1)} = 1. \tag{4.29}$$

If $(z_3+u_3)/(z_2+1) = u_3 - 1$, then Equation (4.29) becomes

$$\frac{(z_2+1)((z_2+1)\hat{\tau}_1+1)}{(z_2+2)(z_2\hat{\tau}_1+1)} = 1,$$

with solution

$$\hat{\tau}_1 = 1.$$

If $(z_3+u_3)/z_2 = u_3 - 1$, then Equation (4.29) becomes

$$\frac{(z_2+1)((z_2+1)\hat{\tau}_2+1)}{(z_2+2)(z_2\hat{\tau}_2+1)} = 1,$$

with solution

$$\hat{\tau}_2 = 1.$$

Otherwise, Equation (4.29) gives

$$\hat{\tau}_1 = \frac{((z_2+2)(z_3+u_3) - (z_2+1)^2(u_3-1))\hat{\tau}_2 + (u_3-1)}{(z_3+u_3)\hat{\tau}_2 + (z_2+1)(z_3+u_3) - z_2(z_2+2)(u_3-1)}. \tag{4.30}$$

Together (4.28) and (4.30) give a quadratic equation in $\hat{\tau}_2$ with roots

$$\hat{\tau}_2 = 1, \frac{-(u_3 - 1)}{(z_3 + u_3)}.$$

The second root is clearly nonpositive, and substituting $\hat{\tau}_2 = 1$ into either (4.28) or (4.30) gives $\hat{\tau}_1 = 1$. Hence, satisfaction of the two 3-cycles previously constructed requires $\hat{\tau}_1 = 1$, unless

$$\frac{z_3 + u_3}{z_1 + 1} = \frac{z_3 + u_3}{z_2} = u_3 - 1\,, \tag{4.31}$$

in which case there is no restriction on the value of $\hat{\tau}_1$. Note that Equation (4.31) cannot be true if $u_3 = 1$.

We can now conclude that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that contains a column with an entry of 1 in a row summing to $z_1 + 1 > 1$ and an entry of 0 in a row summing to $z_2 > 0$ and an entry $u_3 > 0$ in a row summing to $z_3 + u_3 > 1$, must equal 1, unless

$$\frac{z_3 + u_3}{z_1 + 1} = \frac{z_3 + u_3}{z_2} = u_3 - 1\,.$$

Such histograms occur in 3-cycles with histograms of a more arbitrary structure, and it is these that we are now interested in.

**Case (a) $K + 1 = 3$**

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ r-1 \end{pmatrix}, \begin{pmatrix} r-1 \\ 0 \\ 1 \end{pmatrix} \text{and} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1 + r - 1)\Gamma(z_2\hat{\tau}_1)\Gamma((z_3 + r - 1)\hat{\tau}_1 + 1)}{\Gamma((z_1+1)\hat{\tau}_1 + v_1)\Gamma(z_2\hat{\tau}_1 + v_2)\Gamma((z_3 + r - 1)\hat{\tau}_1 + v_3)} \times \\ &\qquad \frac{\Gamma((z_1 + r - 1)\hat{\tau}_2 + v_1)\Gamma(z_2\hat{\tau}_2 + v_2)\Gamma((z_3+1)\hat{\tau}_2 + v_3)}{\Gamma((z_1 + r - 1)\hat{\tau}_2 + 1)\Gamma(z_2\hat{\tau}_2)\Gamma((z_3+1)\hat{\tau}_2 + r - 1)} \times \\ &\qquad \frac{\Gamma((z_1 + v_1)\hat{\tau}_3 + 1)\Gamma((z_3 + v_3)\hat{\tau}_3 + r - 1)}{\Gamma((z_1 + v_1)\hat{\tau}_3 + r - 1)\Gamma((z_3 + v_3)\hat{\tau}_3 + 1)} \\ = \; & 1\,. \end{aligned} \tag{4.32}$$

Assuming that

$$\frac{z_3 + r - 1}{z_1 + 1} = \frac{z_3 + r - 1}{z_2} = r - 2 \tag{4.33}$$

and

$$\frac{z_1+r-1}{z_3+1} = \frac{z_1+r-1}{z_2} = r-2 \tag{4.34}$$

both do *not* hold and that $z_1, z_2, z_3 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.32) gives

$$\frac{\Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3+1)\Gamma((z_3+v_3)\hat{\tau}_3+r-1)}{\Gamma(z_1+1+v_1)\Gamma(z_3+r-1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3+r-1)\Gamma((z_3+v_3)\hat{\tau}_3+1)} = 1\,. \tag{4.35}$$

Equation (4.35) may be rewritten

$$\begin{aligned} & \Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)((z_3+v_3)\hat{\tau}_3+r-2)\cdots((z_3+v_3)\hat{\tau}_3+1) \\ = \; & \Gamma(z_1+1+v_1)\Gamma(z_3+r-1+v_3)((z_1+v_1)\hat{\tau}_3+r-2)\cdots((z_1+v_1)\hat{\tau}_3+1)\,. \end{aligned} \tag{4.36}$$

Consider the ratio of the coefficient of $\hat{\tau}_3^p$ on the left-hand side of (4.36) to the coefficient of $\hat{\tau}_3^p$ on the right-hand side, for $p = 0, \ldots, r-2$. Providing that $z_1 + v_1 \neq z_3 + v_3$, this ratio,

$$\frac{\Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)}{\Gamma(z_1+1+v_1)\Gamma(z_3+r-1+v_3)}\left(\frac{z_3+v_3}{z_1+v_1}\right)^p,$$

when considered as a function of $p$, is monotone and therefore crosses the value 1 at most once. Hence the polynomial in $\hat{\tau}_3$ formed by the difference in the left and right-hand sides of (4.36) has coefficients that change in sign at most once. Theorems 4.3.4 and 4.3.5 then imply that $\hat{\tau}_3 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_3 = 1$ is required to satisfy the current 3-cycle when $z_1 + v_1 \neq z_3 + v_3$.

Suppose Equations (4.33) and (4.34) *do* both hold. It follows that $z_1 + 1 = z_2 = z_3 + 1$. Then $r \neq 3$, because

$$\frac{z_1+2}{z_1+1} \neq 1\,,$$

for any $z_1$. For $r \geq 4$,

$$\begin{aligned} \frac{z_1+r-1}{z_1+1} &= r-2 \\ \Rightarrow z_1 &= \frac{1}{r-3}\,, \end{aligned}$$

which is only possible if, in fact, $r = 4$ and then $z_1 = 1$, $z_2 = 2$, $z_3 = 1$ and $N = 2$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$. Then

$$\frac{1+3}{2+1} \neq \frac{1+3}{1+0} \neq 2$$

and

$$\frac{2+3}{1+1} \neq \frac{2+3}{1+0} \neq 2$$

so that the above approach gives $\hat{\tau}_3 = 1$, unless $2 + v_2 = 1 + v_3$, *i.e.*, unless type $m$ is

$$\begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} \text{ or } \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$$

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$. Then

$$\frac{1+3}{2+1} \neq \frac{1+3}{1+0} \neq 2$$

and

$$\frac{2+3}{1+1} \neq \frac{2+3}{1+0} \neq 2$$

so that the above approach gives $\hat{\tau}_3 = 1$, unless $1 + v_1 = 2 + v_2$, *i.e.*, unless type $m$ is

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} \text{ or } \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Hence one of these last two 3-cycles is guaranteed to give $\hat{\tau}_3 = 1$.

Suppose Equation (4.33) holds but Equation (4.34) does not. Then it is not true that

$$\frac{z_3+r-1}{z_2+1} = \frac{z_3+r-1}{z_1} = r-2\,.$$

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 1 \\ r-1 \end{pmatrix}, \begin{pmatrix} r-1 \\ 0 \\ 1 \end{pmatrix} \text{and} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma(z_1\hat{\tau}_1+r-1)\Gamma((z_2+1)\hat{\tau}_1)\Gamma((z_3+r-1)\hat{\tau}_1+1)}{\Gamma(z_1\hat{\tau}_1+v_1)\Gamma((z_2+1)\hat{\tau}_1+v_2)\Gamma((z_3+r-1)\hat{\tau}_1+v_3)} \times \\ &\qquad \frac{\Gamma((z_1+r-1)\hat{\tau}_2+v_1)\Gamma(z_2\hat{\tau}_2+v_2)\Gamma((z_3+1)\hat{\tau}_2+v_3)}{\Gamma((z_1+r-1)\hat{\tau}_2)\Gamma(z_2\hat{\tau}_2+1)\Gamma((z_3+1)\hat{\tau}_2+r-1)} \times \\ &\qquad \frac{\Gamma((z_1+v_1)\hat{\tau}_3)(z_2+v_2)\hat{\tau}_3\Gamma((z_3+v_3)\hat{\tau}_3+r-1)}{\Gamma((z_1+v_1)\hat{\tau}_3+r-1)\Gamma((z_3+v_3)\hat{\tau}_3+1)} \\ = \;& 1\,. \end{aligned} \tag{4.37}$$

Assuming that $z_1, z_2, z_3 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.37) gives

$$\frac{\Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3)\hat{\tau}_3\Gamma((z_3+v_3)\hat{\tau}_3+r-1)}{\Gamma(z_1+v_1)\Gamma(z_3+r-1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3+r-1)\Gamma((z_3+v_3)\hat{\tau}_3+1)} = 1$$

or

$$\frac{\Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3+1)\Gamma((z_3+v_3)\hat{\tau}_3+r-1)}{\Gamma(z_1+1+v_1)\Gamma(z_3+r-1+v_3)\Gamma((z_1+v_1)\hat{\tau}_3+r-1)\Gamma((z_3+v_3)\hat{\tau}_3+1)} = 1\,. \tag{4.38}$$

Equation (4.38) may be rewritten

$$\begin{aligned} &\Gamma(z_1+r-1+v_1)\Gamma(z_3+1+v_3)((z_3+v_3)\hat{\tau}_3+r-2)\cdots((z_3+v_3)\hat{\tau}_3+1) \\ = \;&\Gamma(z_1+1+v_1)\Gamma(z_3+r-1+v_3)((z_1+v_1)\hat{\tau}_3+r-2)\cdots((z_1+v_1)\hat{\tau}_3+1)\,. \end{aligned} \tag{4.39}$$

Equation (4.39) is identical to Equation (4.36), again leading to the conclusion that $\hat{\tau}_3 = 1$ is required to satisfy the current 3-cycle when $z_1 + v_1 \neq z_3 + v_3$.

Suppose Equation (4.33) does not hold but Equation (4.34) does. An argument similar to the immediately preceding one with types $k$, $l$ and $m$ taken to be

$$\begin{pmatrix} 1 \\ 0 \\ r-1 \end{pmatrix}, \begin{pmatrix} r-1 \\ 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$, shows that $\hat{\tau}_3 = 1$ is required to satisfy this 3-cycle when $z_1 + v_1 \neq z_3 + v_3$.

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that is made up of a positive sub-histogram with row sums $z_1$, $z_2$ and $z_3$ plus a type

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

where $z_1+v_1 \neq z_3+v_3$, must equal 1 to satisfy certain 3-cycles. However, interchanging the roles of some of the categories in types $k$ and $l$ of the above 3-cycles indicates that the only histograms not determined in fact have $z_1 + v_1 = z_2 + v_2 = z_3 + v_3$, *i.e.*, identical row sums. This case will now be considered.

First suppose $r = 3$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

respectively. Assume $z_1, z_2, z_3 > 0$. Then it has been proven that $\hat{\tau}_1 = 1$ unless

$$\frac{z_3+2}{z_1+1} = \frac{z_3+2}{z_2} = 1,$$

that is, unless $z_1 + 1 = z_2 = z_3 + 2$. It has been proven that $\hat{\tau}_2 = 1$ unless $z_1 = z_2 = z_3 + 3$, and that $\hat{\tau}_3 = 1$ unless $z_1 + 1 = z_2 + 1 = z_3 + 1$. Noting that at most one of these three possibilities is true, two members of the set $\{\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3\}$ must equal 1. For this choice of types, (4.26) simplifies to

$$\frac{((z_3+2)\hat{\tau}_1+2)((z_3+2)\hat{\tau}_1+1)}{(z_1+1)\hat{\tau}_1 z_2 \hat{\tau}_1} \frac{z_2\hat{\tau}_2}{((z_3+3)\hat{\tau}_2+1)} \frac{(z_1+1)\hat{\tau}_3}{((z_3+1)\hat{\tau}_3+2)}$$

or

$$\frac{((z_3+2)\hat{\tau}_1+2)((z_3+2)\hat{\tau}_1+1)\hat{\tau}_2\hat{\tau}_3}{\hat{\tau}_1^2((z_3+3)\hat{\tau}_2+1)((z_3+1)\hat{\tau}_3+2)}. \tag{4.40}$$

Substituting $\hat{\tau}_1 = \hat{\tau}_2 = 1$ into (4.40) gives

$$\frac{(z_3+3)\hat{\tau}_3}{((z_3+1)\hat{\tau}_3+2)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,.$$

Substituting $\hat{\tau}_1 = \hat{\tau}_3 = 1$ into (4.40) gives

$$\frac{(z_3+4)\hat{\tau}_2}{((z_3+3)\hat{\tau}_2+1)} = 1\,,$$

with solution

$$\hat{\tau}_2 = 1\,.$$

Substituting $\hat{\tau}_2 = \hat{\tau}_3 = 1$ into (4.40) gives

$$\frac{((z_3+2)\hat{\tau}_1+2)((z_3+2)\hat{\tau}_1+1)}{(z_3+4)(z_3+3)\hat{\tau}_1^2} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1, \frac{-2}{3z_3+8}\,,$$

the second of these values clearly being negative. For $r = 3$ any type is a permutation of one of types $k$, $l$ and $m$. Therefore a suitable rearrangement of the category ordering in this 3-cycle will produce another 3-cycle that gives the required result.

Now assume $r \geq 4$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ r-1 \end{pmatrix}, \begin{pmatrix} r-2 \\ 0 \\ 2 \end{pmatrix} \text{and} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$ and $z_1 + v_1 = z_2 + v_2 = z_3 + v_3$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1+r-2)\Gamma(z_2\hat{\tau}_1)\Gamma((z_3+r-1)\hat{\tau}_1+2)}{\Gamma((z_1+1)\hat{\tau}_1+v_1)\Gamma(z_2\hat{\tau}_1+v_2)\Gamma((z_3+r-1)\hat{\tau}_1+v_3)} \times \\ &\qquad \frac{\Gamma((z_1+r-2)\hat{\tau}_2+v_1)\Gamma(z_2\hat{\tau}_2+v_2)\Gamma((z_3+2)\hat{\tau}_2+v_3)}{\Gamma((z_1+r-2)\hat{\tau}_2+1)\Gamma(z_2\hat{\tau}_2)\Gamma((z_3+2)\hat{\tau}_2+r-1)} \times \\ &\qquad \frac{\Gamma((z_1+v_1)\hat{\tau}_3+1)\Gamma((z_3+v_3)\hat{\tau}_3+r-1)}{\Gamma((z_1+v_1)\hat{\tau}_3+r-2)\Gamma((z_3+v_3)\hat{\tau}_3+2)} \\ = \;& 1\,. \end{aligned} \tag{4.41}$$

Assuming that Equation (4.33) and $z_1 + r - 2 = z_2 = z_3 + 2$ both do *not* hold and that $z_1, z_2, z_3 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.41) gives

$$\frac{\Gamma(z_1 + r - 2 + v_1)\Gamma(z_3 + 2 + v_3)\Gamma((z_1 + v_1)\hat{\tau}_3 + 1)\Gamma((z_3 + v_3)\hat{\tau}_3 + r - 1)}{\Gamma(z_1 + 1 + v_1)\Gamma(z_3 + r - 1 + v_3)\Gamma((z_1 + v_1)\hat{\tau}_3 + r - 2)\Gamma((z_3 + v_3)\hat{\tau}_3 + 2)} = 1$$

or

$$\frac{(z_1 + 1 + v_1)((z_1 + v_1)\hat{\tau}_3 + r - 2)}{(z_1 + r - 2 + v_1)((z_1 + v_1)\hat{\tau}_3 + 1)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,.$$

Suppose Equation (4.33) holds and $z_1 + r - 2 = z_2 = z_3 + 2$. It follows that $r = 3$, a case which has already been dealt with.

Suppose Equation (4.33) holds but $z_1 + r - 2 = z_2 = z_3 + 2$ does not. Then it is not true that

$$\frac{z_3 + r - 1}{z_2 + 1} = \frac{z_3 + r - 1}{z_1} = r - 2\,.$$

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 1 \\ r-1 \end{pmatrix}, \begin{pmatrix} r-2 \\ 0 \\ 2 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

respectively, where $v_3 = r - v_1 - v_2$ and $z_1 + v_1 = z_2 + v_2 = z_3 + v_3$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma(z_1\hat{\tau}_1 + r - 2)\Gamma((z_2 + 1)\hat{\tau}_1)\Gamma((z_3 + r - 1)\hat{\tau}_1 + 2)}{\Gamma(z_1\hat{\tau}_1 + v_1)\Gamma((z_2 + 1)\hat{\tau}_1 + v_2)\Gamma((z_3 + r - 1)\hat{\tau}_1 + v_3)} \times \\ &\qquad \frac{\Gamma((z_1 + r - 2)\hat{\tau}_2 + v_1)\Gamma(z_2\hat{\tau}_2 + v_2)\Gamma((z_3 + 2)\hat{\tau}_2 + v_3)}{\Gamma((z_1 + r - 2)\hat{\tau}_2)\Gamma(z_2\hat{\tau}_2 + 1)\Gamma((z_3 + 2)\hat{\tau}_2 + r - 1)} \times \\ &\qquad \frac{\Gamma((z_1 + v_1)\hat{\tau}_3)(z_2 + v_2)\hat{\tau}_3\Gamma((z_3 + v_3)\hat{\tau}_3 + r - 1)}{\Gamma((z_1 + v_1)\hat{\tau}_3 + r - 2)\Gamma((z_3 + v_3)\hat{\tau}_3 + 2)} \\ = \; & 1\,. \end{aligned} \tag{4.42}$$

Assuming that $z_1, z_2, z_3 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.42) gives

$$\frac{\Gamma(z_1 + r - 2 + v_1)\Gamma(z_3 + 2 + v_3)\Gamma((z_1 + v_1)\hat{\tau}_3)\hat{\tau}_3\Gamma((z_3 + v_3)\hat{\tau}_3 + r - 1)}{\Gamma(z_1 + v_1)\Gamma(z_3 + r - 1 + v_3)\Gamma((z_1 + v_1)\hat{\tau}_3 + r - 2)\Gamma((z_3 + v_3)\hat{\tau}_3 + 2)} = 1$$

or

$$\frac{(z_1+1+v_1)((z_1+v_1)\hat{\tau}_3+r-2)}{(z_1+r-2+v_1)((z_1+v_1)\hat{\tau}_3+1)}=1\,,$$

with solution

$$\hat{\tau}_3=1\,.$$

Suppose Equation (4.33) does not hold but $z_1+r-2=z_2=z_3+2$. Then it is not true that $z_1+r-2=z_2+1=z_3+1$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix}1\\0\\r-1\end{pmatrix},\begin{pmatrix}r-2\\1\\1\end{pmatrix}\text{ and }\begin{pmatrix}r-2\\0\\2\end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned}&\frac{z_2\hat{\tau}_1}{((z_3+r-1)\hat{\tau}_1+1)}\frac{\Gamma((z_1+r-2)\hat{\tau}_2+r-2)\Gamma((z_3+1)\hat{\tau}_2+2)}{\Gamma((z_1+r-2)\hat{\tau}_2+1)\Gamma((z_3+1)\hat{\tau}_2+r-1)}\times\\&\qquad\frac{\Gamma((z_1+r-2)\hat{\tau}_3+1)\Gamma((z_3+2)\hat{\tau}_3+r-1)}{\Gamma((z_1+r-2)\hat{\tau}_3+r-2)z_2\hat{\tau}_3\Gamma((z_3+2)\hat{\tau}_3+1)}\\=\;&1\,.\end{aligned}\tag{4.43}$$

Assuming that $z_1,z_2,z_3>0$, we have already shown that $\hat{\tau}_1=\hat{\tau}_2=1$. Substituting these values into (4.43) gives

$$\frac{\Gamma(z_1+2r-4)\Gamma(z_3+3)\Gamma((z_1+r-2)\hat{\tau}_3+1)\Gamma((z_3+2)\hat{\tau}_3+r-1)}{(z_3+r)\Gamma(z_1+r-1)\Gamma(z_3+r)\Gamma((z_1+r-2)\hat{\tau}_3+r-2)\hat{\tau}_3\Gamma((z_3+2)\hat{\tau}_3+1)}=1$$

or

$$\frac{((z_1+r-2)\hat{\tau}_3+r-2)}{(z_1+2r-4)\hat{\tau}_3}=1\,,$$

with solution

$$\hat{\tau}_3=1\,.$$

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that contains a positive sub-histogram must equal 1 to satisfy certain 3-cycles.

**Case (b) $K+1=4$**

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix}1\\0\\1\\r-2\end{pmatrix},\begin{pmatrix}r-2\\1\\0\\1\end{pmatrix}\text{ and }\begin{pmatrix}v_1\\v_2\\v_3\\v_4\end{pmatrix},$$

respectively, where $v_4 = r - v_1 - v_2 - v_3$. In this case, (4.26) simplifies to

$$\begin{aligned}
&\frac{\Gamma((z_1+1)\hat{\tau}_1 + r - 2)\Gamma(z_2\hat{\tau}_1 + 1)\Gamma((z_3+1)\hat{\tau}_1)\Gamma((z_4+r-2)\hat{\tau}_1 + 1)}{\Gamma((z_1+1)\hat{\tau}_1 + v_1)\Gamma(z_2\hat{\tau}_1 + v_2)\Gamma((z_3+1)\hat{\tau}_1 + v_3)\Gamma((z_4+r-2)\hat{\tau}_1 + v_4)} \times \\
&\quad \frac{\Gamma((z_1+r-2)\hat{\tau}_2 + v_1)\Gamma((z_2+1)\hat{\tau}_2 + v_2)\Gamma(z_3\hat{\tau}_2 + v_3)\Gamma((z_4+1)\hat{\tau}_2 + v_4)}{\Gamma((z_1+r-2)\hat{\tau}_2 + 1)\Gamma((z_2+1)\hat{\tau}_2)\Gamma(z_3\hat{\tau}_2 + 1)\Gamma((z_4+1)\hat{\tau}_2 + r - 2)} \times \\
&\quad \frac{\Gamma((z_1+v_1)\hat{\tau}_3 + 1)(z_3+v_3)\hat{\tau}_3\Gamma((z_4+v_4)\hat{\tau}_3 + r - 2)}{\Gamma((z_1+v_1)\hat{\tau}_3 + r - 2)(z_2+v_2)\hat{\tau}_3\Gamma((z_4+v_4)\hat{\tau}_3 + 1)} \\
= \; & 1 . \qquad (4.44)
\end{aligned}$$

Assuming that $z_1, z_2, z_3, z_4 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.44) gives

$$\frac{\Gamma(z_1 + r - 2 + v_1)\Gamma(z_4 + 1 + v_4)\Gamma((z_1+v_1)\hat{\tau}_3 + 1)\Gamma((z_4+v_4)\hat{\tau}_3 + r - 2)}{\Gamma(z_1 + 1 + v_1)\Gamma(z_4 + r - 2 + v_4)\Gamma((z_1+v_1)\hat{\tau}_3 + r - 2)\Gamma((z_4+v_4)\hat{\tau}_3 + 1)} = 1 . \qquad (4.45)$$

Equation (4.45) may be rewritten

$$\begin{aligned}
& \Gamma(z_1 + r - 2 + v_1)\Gamma(z_4 + 1 + v_4)((z_4+v_4)\hat{\tau}_3 + r - 3)\cdots((z_4+v_4)\hat{\tau}_3 + 1) \\
= \; & \Gamma(z_1 + 1 + v_1)\Gamma(z_4 + r - 2 + v_4)((z_1+v_1)\hat{\tau}_3 + r - 3)\cdots((z_1+v_1)\hat{\tau}_3 + 1) . \qquad (4.46)
\end{aligned}$$

Consider the ratio of the coefficient of $\hat{\tau}_3^p$ on the left-hand side of (4.46) to the coefficient of $\hat{\tau}_3^p$ on the right-hand side, for $p = 0, \ldots, r-3$. Providing that $z_1 + v_1 \neq z_4 + v_4$, this ratio,

$$\frac{\Gamma(z_1 + r - 2 + v_1)\Gamma(z_4 + 1 + v_4)}{\Gamma(z_1 + 1 + v_1)\Gamma(z_4 + r - 2 + v_4)} \left(\frac{z_4 + v_4}{z_1 + v_1}\right)^p ,$$

when considered as a function of $p$, is monotone and therefore crosses the value 1 at most once. Hence the polynomial in $\hat{\tau}_3$ formed by the difference in the left and right-hand sides of (4.46) has coefficients that change in sign at most once. Theorems 4.3.4 and 4.3.5 then imply that $\hat{\tau}_3 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_3 = 1$ is required to satisfy the current 3-cycle when $z_1 + v_1 \neq z_4 + v_4$.

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that is made up of a positive sub-histogram with row sums $z_1$, $z_2$, $z_3$ and $z_4$ plus a

type

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix},$$

where $z_1 + v_1 \neq z_4 + v_4$, must equal 1 to satisfy certain 3-cycles. The case where $z_1 + v_1 = z_4 + v_4$ will now be considered.

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ r-2 \end{pmatrix}, \begin{pmatrix} r-1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix},$$

respectively, where $v_4 = r - v_1 - v_2 - v_3$ and $z_1 + v_1 = z_4 + v_4$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1 + r - 1)\Gamma(z_2\hat{\tau}_1)\Gamma((z_3+1)\hat{\tau}_1 + 1)\Gamma((z_4+r-2)\hat{\tau}_1)}{\Gamma((z_1+1)\hat{\tau}_1 + v_1)\Gamma(z_2\hat{\tau}_1 + v_2)\Gamma((z_3+1)\hat{\tau}_1 + v_3)\Gamma((z_4+r-2)\hat{\tau}_1 + v_4)} \times \\ &\quad \frac{\Gamma((z_1+r-1)\hat{\tau}_2 + v_1)\Gamma(z_2\hat{\tau}_2 + v_2)\Gamma((z_3+1)\hat{\tau}_2 + v_3)\Gamma(z_4\hat{\tau}_2 + v_4)}{\Gamma((z_1+r-1)\hat{\tau}_2 + 1)\Gamma(z_2\hat{\tau}_2)\Gamma((z_3+1)\hat{\tau}_2 + 1)\Gamma(z_4\hat{\tau}_2 + r - 2)} \times \\ &\quad \frac{\Gamma((z_1+v_1)\hat{\tau}_3 + 1)\Gamma((z_4+v_4)\hat{\tau}_3 + r - 2)}{\Gamma((z_1+v_1)\hat{\tau}_3 + r - 1)\Gamma((z_4+v_4)\hat{\tau}_3)} \\ = \; & 1\,. \end{aligned} \tag{4.47}$$

Assuming that

$$\frac{z_1 + r - 1}{z_3 + 1} = \frac{z_1 + r - 1}{z_2} = r - 2 \tag{4.48}$$

does *not* hold and that $z_1, z_2, z_3, z_4 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.47) gives

$$\frac{\Gamma(z_1 + r - 1 + v_1)\Gamma(z_4 + v_4)\Gamma((z_1+v_1)\hat{\tau}_3 + 1)\Gamma((z_4+v_4)\hat{\tau}_3 + r - 2)}{\Gamma(z_1 + 1 + v_1)\Gamma(z_4 + r - 2 + v_4)\Gamma((z_1+v_1)\hat{\tau}_3 + r - 1)\Gamma((z_4+v_4)\hat{\tau}_3)} = 1$$

or

$$\frac{(z_1 + r - 2 + v_1)(z_1 + v_1)\hat{\tau}_3}{(z_1 + v_1)((z_1+v_1)\hat{\tau}_3 + r - 2)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,.$$

Suppose Equation (4.48) holds. Then it is not true that

$$\frac{z_1+r-1}{z_2+1} = \frac{z_1+r-1}{z_3} = r-2\,.$$

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ r-2 \end{pmatrix}, \begin{pmatrix} r-1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix},$$

respectively, where $v_4 = r - v_1 - v_2 - v_3$ and $z_1 + v_1 = z_4 + v_4$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1+r-1)\Gamma(z_2\hat{\tau}_1+1)\Gamma((z_3+1)\hat{\tau}_1)\Gamma((z_4+r-2)\hat{\tau}_1)}{\Gamma((z_1+1)\hat{\tau}_1+v_1)\Gamma(z_2\hat{\tau}_1+v_2)\Gamma((z_3+1)\hat{\tau}_1+v_3)\Gamma((z_4+r-2)\hat{\tau}_1+v_4)} \times \\ &\quad \frac{\Gamma((z_1+r-1)\hat{\tau}_2+v_1)\Gamma((z_2+1)\hat{\tau}_2+v_2)\Gamma(z_3\hat{\tau}_2+v_3)\Gamma(z_4\hat{\tau}_2+v_4)}{\Gamma((z_1+r-1)\hat{\tau}_2+1)\Gamma((z_2+1)\hat{\tau}_2)\Gamma(z_3\hat{\tau}_2+1)\Gamma(z_4\hat{\tau}_2+r-2)} \times \\ &\quad \frac{\Gamma((z_1+v_1)\hat{\tau}_3+1)(z_3+v_3)\hat{\tau}_3\Gamma((z_4+v_4)\hat{\tau}_3+r-2)}{\Gamma((z_1+v_1)\hat{\tau}_3+r-1)(z_2+v_2)\hat{\tau}_3\Gamma((z_4+v_4)\hat{\tau}_3)} \\ = \;& 1\,. \end{aligned} \tag{4.49}$$

Assuming that $z_1, z_2, z_3, z_4 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.49) gives

$$\frac{\Gamma(z_1+r-1+v_1)\Gamma(z_4+v_4)\Gamma((z_1+v_1)\hat{\tau}_3+1)\Gamma((z_4+v_4)\hat{\tau}_3+r-2)}{\Gamma(z_1+1+v_1)\Gamma(z_4+r-2+v_4)\Gamma((z_1+v_1)\hat{\tau}_3+r-1)\Gamma((z_4+v_4)\hat{\tau}_3)} = 1$$

or

$$\frac{(z_1+r-2+v_1)((z_1+v_1)\hat{\tau}_3}{(z_1+v_1)((z_1+v_1)\hat{\tau}_3+r-2)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,.$$

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that contains a positive sub-histogram must equal 1 to satisfy certain 3-cycles.

**Case (c) $K+1 \geq 5$**

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \\ u_5 \\ \vdots \\ u_{K+1} \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \\ u_5 \\ \vdots \\ u_{K+1} \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ \vdots \\ v_{K+1} \end{pmatrix},$$

respectively, where $u_{K+1} = r - 4 - \sum_{j=5}^{K} u_j$ ($\sum_{j=5}^{K} u_j = 0$ if $K = 4$) and $v_{K+1} = r - \sum_{j=1}^{K} v_j$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1+2)\Gamma(z_2\hat{\tau}_1+1)\Gamma((z_3+1)\hat{\tau}_1)\Gamma((z_4+2)\hat{\tau}_1+1)}{\Gamma((z_1+1)\hat{\tau}_1+v_1)\Gamma(z_2\hat{\tau}_1+v_2)\Gamma((z_3+1)\hat{\tau}_1+v_3)\Gamma((z_4+2)\hat{\tau}_1+v_4)} \times \\ &\quad \frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+v_j)} \times \\ &\quad \frac{\Gamma((z_1+2)\hat{\tau}_2+v_1)\Gamma((z_2+1)\hat{\tau}_2+v_2)\Gamma(z_3\hat{\tau}_2+v_3)\Gamma((z_4+1)\hat{\tau}_2+v_4)}{\Gamma((z_1+2)\hat{\tau}_2+1)\Gamma((z_2+1)\hat{\tau}_2)\Gamma(z_3\hat{\tau}_2+1)\Gamma((z_4+1)\hat{\tau}_2+2)} \times \\ &\quad \frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_2+v_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)} \times \\ &\quad \frac{(z_3+v_3)\hat{\tau}_3((z_4+v_4)\hat{\tau}_3+1)}{((z_1+v_1)\hat{\tau}_3+1)(z_2+v_2)\hat{\tau}_3} \\ = \; &1\,. \end{aligned} \tag{4.50}$$

Assuming that $z_1, z_2, z_3, z_4 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.50) gives

$$\frac{(z_1+1+v_1)((z_4+v_4)\hat{\tau}_3+1)}{(z_4+1+v_4)((z_1+v_1)\hat{\tau}_3+1)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,,$$

providing $z_1 + v_1 \neq z_4 + v_4$.

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that is made up of a sub-histogram containing at least four positive row sums $z_1$,

$z_2$, $z_3$ and $z_4$ plus a type

$$\begin{pmatrix} v_1 \\ \vdots \\ v_{K+1} \end{pmatrix},$$

where $z_1 + v_1 \neq z_4 + v_4$, must equal 1 to satisfy certain 3-cycles. The case where $z_1 + v_1 = z_4 + v_4$ will now be considered.

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 2 \\ u_5 \\ \vdots \\ u_{K+1} \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \\ 0 \\ u_5 \\ \vdots \\ u_{K+1} \end{pmatrix} \text{ and } \begin{pmatrix} v_1 \\ \vdots \\ v_{K+1} \end{pmatrix},$$

respectively, where $u_{K+1} = r-4-\sum_{j=5}^{K} u_j$ ($\sum_{j=5}^{K} u_j = 0$ if $K = 4$), $v_{K+1} = r-\sum_{j=1}^{K} v_j$ and $z_1 + v_1 = z_4 + v_4$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{\Gamma((z_1+1)\hat{\tau}_1+3)\Gamma(z_2\hat{\tau}_1)\Gamma((z_3+1)\hat{\tau}_1+1)\Gamma((z_4+2)\hat{\tau}_1)}{\Gamma((z_1+1)\hat{\tau}_1+v_1)\Gamma(z_2\hat{\tau}_1+v_2)\Gamma((z_3+1)\hat{\tau}_1+v_3)\Gamma((z_4+2)\hat{\tau}_1+v_4)} \times \\ &\quad \frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+v_j)} \times \\ &\quad \frac{\Gamma((z_1+3)\hat{\tau}_2+v_1)\Gamma(z_2\hat{\tau}_2+v_2)\Gamma((z_3+1)\hat{\tau}_2+v_3)\Gamma(z_4\hat{\tau}_2+v_4)}{\Gamma((z_1+3)\hat{\tau}_2+1)\Gamma(z_2\hat{\tau}_2)\Gamma((z_3+1)\hat{\tau}_2+1)\Gamma(z_4\hat{\tau}_2+2)} \times \\ &\quad \frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_2+v_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)} \times \\ &\quad \frac{((z_4+v_4)\hat{\tau}_3+1)(z_4+v_4)\hat{\tau}_3}{((z_1+v_1)\hat{\tau}_3+2)((z_1+v_1)\hat{\tau}_3+1)} \\ = \ &1\,. \end{aligned} \tag{4.51}$$

Assuming that

$$\frac{z_1+3}{z_3+1} = \frac{z_1+3}{z_2} = 2 \tag{4.52}$$

does *not* hold and $z_1, z_2, z_3, z_4 > 0$, we have already shown that $\hat{\tau}_1 = \hat{\tau}_2 = 1$. Substituting these values into (4.51) gives

$$\frac{(z_1+2+v_1)(z_1+1+v_1)((z_4+v_4)\hat{\tau}_3+1)\hat{\tau}_3}{(z_4+1+v_4)((z_1+v_1)\hat{\tau}_3+2)((z_1+v_1)\hat{\tau}_3+1)} = 1$$

or

$$\frac{(z_1+2+v_1)\hat{\tau}_3}{((z_1+v_1)\hat{\tau}_3+2)}=1\,,$$

with solution

$$\hat{\tau}_3=1\,.$$

Suppose Equation (4.52) holds. Then it is not true that

$$\frac{z_1+3}{z_2+1}=\frac{z_1+3}{z_3}=2\,.$$

Let types $k$, $l$ and $m$ be

$$\begin{pmatrix}1\\0\\1\\2\\u_5\\\vdots\\u_{K+1}\end{pmatrix},\begin{pmatrix}3\\1\\0\\0\\u_5\\\vdots\\u_{K+1}\end{pmatrix}\text{ and }\begin{pmatrix}v_1\\\vdots\\v_{K+1}\end{pmatrix},$$

respectively, where $u_{K+1}=r-4-\sum_{j=5}^{K}u_j$ ($\sum_{j=5}^{K}u_j=0$ if $K=4$), $v_{K+1}=r-\sum_{j=1}^{K}v_j$ and $z_1+v_1=z_4+v_4$. Then (4.26) simplifies to

$$\begin{aligned}&\frac{\Gamma((z_1+1)\hat{\tau}_1+3)\Gamma(z_2\hat{\tau}_1+1)\Gamma((z_3+1)\hat{\tau}_1)\Gamma((z_4+2)\hat{\tau}_1)}{\Gamma((z_1+1)\hat{\tau}_1+v_1)\Gamma(z_2\hat{\tau}_1+v_2)\Gamma((z_3+1)\hat{\tau}_1+v_3)\Gamma((z_4+2)\hat{\tau}_1+v_4)}\times\\&\quad\frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+v_j)}\times\\&\quad\frac{\Gamma((z_1+3)\hat{\tau}_2+v_1)\Gamma((z_2+1)\hat{\tau}_2+v_2)\Gamma(z_3\hat{\tau}_2+v_3)\Gamma(z_4\hat{\tau}_2+v_4)}{\Gamma((z_1+3)\hat{\tau}_2+1)\Gamma((z_2+1)\hat{\tau}_2)\Gamma(z_3\hat{\tau}_2+1)\Gamma(z_4\hat{\tau}_2+2)}\times\\&\quad\frac{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_2+v_j)}{\prod_{j=5}^{K+1}\Gamma((z_j+u_j)\hat{\tau}_1+u_j)}\times\\&\quad\frac{(z_3+v_3)\hat{\tau}_3((z_4+v_4)\hat{\tau}_3+1)(z_4+v_4)\hat{\tau}_3}{((z_1+v_1)\hat{\tau}_3+2)((z_1+v_1)\hat{\tau}_3+1)(z_2+v_2)\hat{\tau}_3}\\=\;&1\,.\end{aligned}\tag{4.53}$$

Assuming that $z_1,z_2,z_3,z_4>0$, we have already shown that $\hat{\tau}_1=\hat{\tau}_2=1$. Substituting these values into (4.53) gives

$$\frac{(z_1+2+v_1)(z_1+1+v_1)((z_4+v_4)\hat{\tau}_3+1)\hat{\tau}_3}{(z_4+1+v_4)((z_1+v_1)\hat{\tau}_3+2)((z_1+v_1)\hat{\tau}_3+1)}=1$$

or

$$\frac{(z_1 + 2 + v_1)\hat{\tau}_3}{((z_1 + v_1)\hat{\tau}_3 + 2)} = 1\,,$$

with solution

$$\hat{\tau}_3 = 1\,.$$

Hence we have shown that the $\hat{\tau}$ variable associated with any histogram in $\mathcal{H}_P$ that contains a sub-histogram with at least four positive rows must equal 1 to satisfy certain 3-cycles.

It remains to deal with the special cases not covered by the proof so far.

**Special Cases (a) $K + 1 = 3$**

It is necessary to characterise those histograms in $\mathcal{H}_P$ that do not contain a positive sub-histogram. Consider such a histogram. For $i = 1, \ldots, N$, suppose that the sub-histogram obtained by removing column $i$ has $g_i$ zero rows. Then for $i = 1, \ldots, N$, $1 \leq g_i \leq 2$ and column $i$ is the only one with positive entries in those $g_i$ rows. Fix $I \in \{1, \ldots, N\}$. Then the sub-histogram obtained by removing column $I$ must have at least $\sum_{\substack{i=1 \\ i \neq I}}^{N} g_i$ positive rows. Hence

$$N - 1 \leq \sum_{\substack{i=1 \\ i \neq I}}^{N} g_i \leq 2\,,$$

which implies that $N \leq 3$.

Assume $N = 2$. It is clear that the original histogram is some category permutation of the form

$$\begin{bmatrix} u & 0 \\ r-u & 0 \\ 0 & r \end{bmatrix} \text{ or } \begin{bmatrix} u & v \\ r-u & 0 \\ 0 & r-v \end{bmatrix}\,,$$

where $u, v \in \{1, \ldots, r-1\}$. The case $N + 1 = 3$, $r = 3$, $K + 1 = 3$ has been dealt with empirically (see Table 4.3) and the theorem holds. Hence, assume $r \geq 4$. Let the sub-histogram be

$$\begin{bmatrix} u \\ r-u \\ 0 \end{bmatrix}\,,$$

where $u \in \{1, \ldots, r-1\}$, and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ r-2 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ r-2 \\ 1 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned} & \frac{\Gamma((r-u)\hat{\tau}_1+1)\Gamma(r\hat{\tau}_1+r-2)}{\Gamma((r-u)\hat{\tau}_1+r-2)\Gamma(r\hat{\tau}_1+1)} \times \\ & \qquad \frac{(u+1)\hat{\tau}_2\Gamma((r-u+1)\hat{\tau}_2+r-2)\Gamma((r-2)\hat{\tau}_2+1)}{\Gamma((r-u+1)\hat{\tau}_2)\Gamma((r-2)\hat{\tau}_2+r)} \times \\ & \qquad \frac{(\hat{\tau}_3+r-1)(\hat{\tau}_3+r-2)}{(u+1)\hat{\tau}_3(2r-u-2)\hat{\tau}_3} \\ = \; & 1\,. \end{aligned} \tag{4.54}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.54) gives

$$\frac{\Gamma(2r-u-1)\Gamma(r-1)r(r-1)\Gamma((r-u)\hat{\tau}_1+1)\Gamma(r\hat{\tau}_1+r-2)}{\Gamma(r-u+1)\Gamma(2r-2)(2r-u-2)\Gamma((r-u)\hat{\tau}_1+r-2)\Gamma(r\hat{\tau}_1+1)} = 1$$

or

$$\frac{\Gamma(2r-u-2)\Gamma(r+1)\Gamma((r-u)\hat{\tau}_1+1)\Gamma(r\hat{\tau}_1+r-2)}{\Gamma(r-u+1)\Gamma(2r-2)\Gamma((r-u)\hat{\tau}_1+r-2)\Gamma(r\hat{\tau}_1+1)} = 1\,. \tag{4.55}$$

Equation (4.55) may be rewritten

$$\begin{aligned} & \Gamma(2r-u-2)\Gamma(r+1)(r\hat{\tau}_1+r-3)\cdots(r\hat{\tau}_1+1) \\ = \; & \Gamma(r-u+1)\Gamma(2r-2)((r-u)\hat{\tau}_1+r-3)\cdots((r-u)\hat{\tau}_1+1)\,. \end{aligned} \tag{4.56}$$

Consider the ratio of the coefficient of $\hat{\tau}_1^p$ on the left-hand side of (4.56) to the coefficient of $\hat{\tau}_1^p$ on the right-hand side, for $p = 0, \ldots, r-3$. Since $r - u \neq u$, this ratio,

$$\frac{\Gamma(2r-u-2)\Gamma(r+1)}{\Gamma(r-u+1)\Gamma(2r-2)}\left(\frac{r}{r-u}\right)^p,$$

when considered as a function of $p$, is monotone and therefore crosses the value 1 at most once. Hence the polynomial in $\hat{\tau}_1$ formed by the difference in the left and right-hand sides of (4.56) has coefficients that change in sign at most once. Theorems 4.3.4 and 4.3.5 then imply that $\hat{\tau}_1 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_1 = 1$ is required to satisfy the current 3-cycle.

Now let types $k$, $l$ and $m$ be

$$\begin{pmatrix} v \\ 0 \\ r-v \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 1 \\ r-2 \end{pmatrix},$$

respectively, where $v \in \{1, \ldots, r-1\}$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{((r-v)\hat{\tau}_1 + r - 1)((r-v)\hat{\tau}_1 + r - 2)}{(u+v)\hat{\tau}_1(r-u)\hat{\tau}_1} \frac{\Gamma(u\hat{\tau}_2 + 1)(r-u)\hat{\tau}_2\Gamma(r\hat{\tau}_2 + r - 2)}{\Gamma(u\hat{\tau}_2 + v)\Gamma(r\hat{\tau}_2 + r - v)} \times \\ &\quad \frac{\Gamma((u+1)\hat{\tau}_3 + v)\Gamma((r-2)\hat{\tau}_3 + r - v)}{\Gamma((u+1)\hat{\tau}_3)\Gamma((r-2)\hat{\tau}_3 + r)} \\ = \; &1 . \end{aligned} \tag{4.57}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.57) gives

$$\frac{((r-v)\hat{\tau}_1 + r - 1)((r-v)\hat{\tau}_1 + r - 2)}{(2r-v-1)(2r-v-2)\hat{\tau}_1^2} = 1 ,$$

with solution

$$\hat{\tau}_1 = 1, \frac{-(r-1)(r-2)}{(r(r-3) + (r-v)(2r-3) + 2)} ,$$

the second of these values clearly being negative.

Assume $N = 3$. It is clear that the original histogram is of the form

$$\begin{bmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \end{bmatrix} .$$

Let the sub-histogram be

$$\begin{bmatrix} r & 0 \\ 0 & r \\ 0 & 0 \end{bmatrix}$$

and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ r-2 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ 0 \\ r-2 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned} & \frac{r\hat{\tau}_1}{(r\hat{\tau}_1+1)} \frac{((r+1)\hat{\tau}_2+1)(r+1)\hat{\tau}_2}{((r-2)\hat{\tau}_2+r-1)((r-2)\hat{\tau}_2+r-2)} \times \\ & \qquad \frac{((r-2)\hat{\tau}_3+r-1)((r-2)\hat{\tau}_3+r-2)}{(r+2)\hat{\tau}_3 r\hat{\tau}_3} \\ = \; & 1 . \end{aligned} \tag{4.58}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.58) gives

$$\frac{(r+1)\hat{\tau}_1}{(r\hat{\tau}_1+1)} = 1 ,$$

with solution

$$\hat{\tau}_1 = 1 .$$

**Special Cases (b) $K+1 \geq 4$**

It is necessary to characterise those histograms in $\mathcal{H}_P$ that do not contain a sub-histogram with at least 4 positive rows. Consider such a histogram. For $i = 1, \ldots, N$, suppose that the sub-histogram obtained by removing column $i$ has $g_i$ zero rows. Then for $i = 1, \ldots, N$, $K-2 \leq g_i \leq K$ and column $i$ is the only one with positive entries in those $g_i$ rows. Fix $I \in \{1, \ldots, N\}$. Then the sub-histogram obtained by removing column $I$ must have at least $\sum_{\substack{i=1 \\ i \neq I}}^{N} g_i$ positive rows. Hence

$$(N-1)(K-2) \leq \sum_{\substack{i=1 \\ i \neq I}}^{N} g_i \leq 3 , \tag{4.59}$$

Assume $K+1 = 4$. Then Equation (4.59) implies that $N \leq 4$.

Assume $N = 2$. It is clear that the original histogram is some category permutation of the form

$$\begin{bmatrix} u_1 & 0 \\ u_2 & 0 \\ u_3 & 0 \\ 0 & r \end{bmatrix}, \begin{bmatrix} u_1 & v \\ u_2 & 0 \\ u_3 & 0 \\ 0 & r-v \end{bmatrix}, \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \\ u_3 & 0 \\ 0 & v_3 \end{bmatrix} \text{ or } \begin{bmatrix} u & 0 \\ r-u & 0 \\ 0 & v \\ 0 & r-v \end{bmatrix},$$

where $u_1, u_2, u_3, v_1, v_2, v_3 \in \{1, \ldots, r-2\}$, $u, v \in \{1, \ldots, r-1\}$, and $u_3 = r - u_1 - u_2$,

$v_3 = r - v_1 - v_2$. Let the sub-histogram be

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 0 \end{bmatrix},$$

where $u_1, u_2, u_3 \in \{1, \ldots, r-2\}$ and $u_3 = r - u_1 - u_2$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix} \text{and} \begin{pmatrix} 1 \\ 1 \\ r-3 \\ 1 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned} & \frac{\Gamma(u_3\hat{\tau}_1 + 1)\Gamma(r\hat{\tau}_1 + r - 3)}{\Gamma(u_3\hat{\tau}_1 + r - 3)\Gamma(r\hat{\tau}_1 + 1)} \times \\ & \qquad \frac{(u_1+1)\hat{\tau}_2(u_2+1)\hat{\tau}_2\Gamma((u_3+1)\hat{\tau}_2 + r - 3)\Gamma((r-3)\hat{\tau}_2 + 1)}{\Gamma((u_3+1)\hat{\tau}_2)\Gamma((r-3)\hat{\tau}_2 + r)} \times \\ & \qquad \frac{\Gamma(\hat{\tau}_3 + r)}{(u_1+1)\hat{\tau}_3(u_2+1)\hat{\tau}_3(u_3 + r - 3)\hat{\tau}_3\Gamma(\hat{\tau}_3 + r - 3)} \\ = \; & 1 . \end{aligned} \tag{4.60}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.60) gives

$$\frac{\Gamma(u_3 + r - 3)\Gamma(r+1)\Gamma(u_3\hat{\tau}_1 + 1)\Gamma(r\hat{\tau}_1 + r - 3)}{\Gamma(u_3+1)\Gamma(2r-3)\Gamma(u_3\hat{\tau}_1 + r - 3)\Gamma(r\hat{\tau}_1 + 1)} = 1 . \tag{4.61}$$

Equation (4.61) may be rewritten

$$\begin{aligned} & \Gamma(u_3 + r - 3)\Gamma(r+1)(r\hat{\tau}_1 + r - 4)\cdots(r\hat{\tau}_1 + 1) \\ = \; & \Gamma(u_3+1)\Gamma(2r-3)(u_3\hat{\tau}_1 + r - 4)\cdots(u_3\hat{\tau}_1 + 1) . \end{aligned} \tag{4.62}$$

Consider the ratio of the coefficient of $\hat{\tau}_1^p$ on the left-hand side of (4.62) to the coefficient of $\hat{\tau}_1^p$ on the right-hand side, for $p = 0, \ldots, r-4$. Since $u_3 \neq r$, this ratio,

$$\frac{\Gamma(u_3 + r - 3)\Gamma(r+1)}{\Gamma(u_3+1)\Gamma(2r-3)} \left(\frac{r}{u_3}\right)^p ,$$

when considered as a function of $p$, is monotone and therefore crosses the value 1 at most once. Hence the polynomial in $\hat{\tau}_1$ formed by the difference in the left

and right-hand sides of (4.62) has coefficients that change in sign at most once. Theorems 4.3.4 and 4.3.5 then imply that $\hat{\tau}_1 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_1 = 1$ is required to satisfy the current 3-cycle.

Now let types $k$, $l$ and $m$ be

$$\begin{pmatrix} v \\ 0 \\ 0 \\ r-v \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix},$$

respectively, where $v \in \{1, \ldots, r-1\}$. Then (4.26) simplifies to

$$\begin{aligned} & \frac{((r-v)\hat{\tau}_1 + r - 1)((r-v)\hat{\tau}_1 + r - 2)((r-v)\hat{\tau}_1 + r - 3)}{(u_1 + v)\hat{\tau}_1 u_2 \hat{\tau}_1 u_3 \hat{\tau}_1} \times \\ & \quad \frac{\Gamma(u_1\hat{\tau}_2 + 1)u_2\hat{\tau}_2 u_3 \hat{\tau}_2 \Gamma(r\hat{\tau}_2 + r - 3)}{\Gamma(u_1\hat{\tau}_2 + v)\Gamma(r\hat{\tau}_2 + r - v)} \times \\ & \quad \frac{\Gamma((u_1+1)\hat{\tau}_3 + v)\Gamma((r-3)\hat{\tau}_3 + r - v)}{\Gamma((u_1+1)\hat{\tau}_3)\Gamma((r-3)\hat{\tau}_3 + r)} \\ = \; & 1 . \end{aligned} \tag{4.63}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.63) gives

$$\frac{((r-v)\hat{\tau}_1 + r - 1)((r-v)\hat{\tau}_1 + r - 2)((r-v)\hat{\tau}_1 + r - 3)}{(2r - v - 1)(2r - v - 2)(2r - v - 3)\hat{\tau}_1^3} = 1 . \tag{4.64}$$

Equation (4.64) may be rewritten

$$\begin{aligned} & ((r-v)\hat{\tau}_1 + r - 1)((r-v)\hat{\tau}_1 + r - 2)((r-v)\hat{\tau}_1 + r - 3) \\ = \; & (2r - v - 1)(2r - v - 2)(2r - v - 3)\hat{\tau}_1^3 . \end{aligned} \tag{4.65}$$

The polynomial in $\hat{\tau}_1$ formed by the difference in the left and right-hand sides of (4.65) has coefficients that change in sign exactly once. Theorem 4.3.5 then implies that $\hat{\tau}_1 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_1 = 1$ is required to satisfy the current 3-cycle.

Now let types $k$, $l$ and $m$ be

$$\begin{pmatrix} v_1 \\ v_2 \\ 0 \\ v_3 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix},$$

respectively, where $v_1, v_2, v_3 \in \{1, \ldots, r-2\}$ and $v_3 = r - v_1 - v_2$. Then (4.26) simplifies to

$$\begin{aligned} & \frac{(v_3\hat{\tau}_1 + r - 1)(v_3\hat{\tau}_1 + r - 2)(v_3\hat{\tau}_1 + r - 3)}{(u_1 + v_1)\hat{\tau}_1(u_2 + v_2)\hat{\tau}_1 u_3\hat{\tau}_1} \times \\ & \quad \frac{\Gamma(u_1\hat{\tau}_2 + 1)\Gamma(u_2\hat{\tau}_2 + 1)u_3\hat{\tau}_2\Gamma(r\hat{\tau}_2 + r - 3)}{\Gamma(u_1\hat{\tau}_2 + v_1)\Gamma(u_2\hat{\tau}_2 + v_2)\Gamma(r\hat{\tau}_2 + v_3)} \times \\ & \quad \frac{\Gamma((u_1 + 1)\hat{\tau}_3 + v_1)\Gamma((u_2 + 1)\hat{\tau}_3 + v_2)\Gamma((r-3)\hat{\tau}_3 + v_3)}{\Gamma((u_1 + 1)\hat{\tau}_3)\Gamma((u_2 + 1)\hat{\tau}_3)\Gamma((r-3)\hat{\tau}_3 + r)} \\ = \; & 1 . \end{aligned} \tag{4.66}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.66) gives

$$\frac{(v_3\hat{\tau}_1 + r - 1)(v_3\hat{\tau}_1 + r - 2)(v_3\hat{\tau}_1 + r - 3)}{(r + v_3 - 1)(r + v_3 - 2)(r + v_3 - 3)\hat{\tau}_1^3} = 1 . \tag{4.67}$$

Equation (4.67) may be rewritten

$$\begin{aligned} & (v_3\hat{\tau}_1 + r - 1)(v_3\hat{\tau}_1 + r - 2)(v_3\hat{\tau}_1 + r - 3) \\ = \; & (r + v_3 - 1)(r + v_3 - 2)(r + v_3 - 3)\hat{\tau}_1^3 . \end{aligned} \tag{4.68}$$

The polynomial in $\hat{\tau}_1$ formed by the difference in the left and right-hand sides of (4.68) has coefficients that change in sign exactly once. Theorem 4.3.5 then implies that $\hat{\tau}_1 = 1$ is the only positive root of this polynomial or, in other words, that $\hat{\tau}_1 = 1$ is required to satisfy the current 3-cycle.

Now let the sub-histogram be

$$\begin{bmatrix} u \\ r - u \\ 0 \\ 0 \end{bmatrix},$$

where $u \in \{1, \ldots, r-1\}$, and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ v \\ r - v \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ r - 3 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ 0 \\ 1 \\ r - 3 \end{pmatrix},$$

respectively, where $v \in \{1, \ldots, r-1\}$. Then (4.26) simplifies to

$$\begin{aligned} & \frac{(r-u)\hat{\tau}_1}{(u\hat{\tau}_1+1)} \frac{((u+1)\hat{\tau}_2+1)(u+1)\hat{\tau}_2\Gamma(\hat{\tau}_2+1)\Gamma((r-3)\hat{\tau}_2+r-3)}{\Gamma(\hat{\tau}_2+v)\Gamma((r-3)\hat{\tau}_2+r-v)} \times \\ & \qquad \frac{\Gamma(\hat{\tau}_3+v)\Gamma((r-3)\hat{\tau}_3+r-v)}{(u+2)\hat{\tau}_3(r-u)\hat{\tau}_3\Gamma(\hat{\tau}_3+1)\Gamma((r-3)\hat{\tau}_3+r-3)} \\ = \; & 1\,. \end{aligned} \tag{4.69}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.69) gives

$$\frac{(u+1)\hat{\tau}_1}{(u\hat{\tau}_1+1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Assume $N = 3$. It is clear that the original histogram is some category permutation of the form

$$\begin{bmatrix} u & 0 & 0 \\ r-u & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \end{bmatrix}, \begin{bmatrix} u & v & 0 \\ r-u & 0 & 0 \\ 0 & r-v & 0 \\ 0 & 0 & r \end{bmatrix} \text{or} \begin{bmatrix} u & v & w \\ r-u & 0 & 0 \\ 0 & r-v & 0 \\ 0 & 0 & r-w \end{bmatrix},$$

where $u, v, w \in \{1, \ldots, r-1\}$. Let the sub-histogram be

$$\begin{bmatrix} u & 0 \\ r-u & 0 \\ 0 & r \\ 0 & 0 \end{bmatrix},$$

where $u \in \{1, \ldots, r-1\}$, and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} r-3 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{and} \begin{pmatrix} r-3 \\ 1 \\ 0 \\ 2 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned} & \frac{r\hat{\tau}_1}{(r\hat{\tau}_1+1)} \frac{\Gamma((u+r-3)\hat{\tau}_2+r-3)(r-u+1)\hat{\tau}_2\Gamma(\hat{\tau}_2+2)}{\Gamma((u+r-3)\hat{\tau}_2)\Gamma(\hat{\tau}_2+r)} \times \\ & \qquad \frac{\Gamma((u+r-3)\hat{\tau}_3)\Gamma(2\hat{\tau}_3+r)}{\Gamma((u+r-3)\hat{\tau}_3+r-3)(r-u+1)\hat{\tau}_3 r\hat{\tau}_3\Gamma(2\hat{\tau}_3+1)} \\ = \; & 1\,. \end{aligned} \tag{4.70}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.70) gives

$$\frac{(1+r)\hat{\tau}_1}{(r\hat{\tau}_1+1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Now let the sub-histogram be

$$\begin{bmatrix} u & v \\ r-u & 0 \\ 0 & r-v \\ 0 & 0 \end{bmatrix},$$

where $u, v \in \{1, \ldots, r-1\}$, and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix} \text{and} \begin{pmatrix} 2 \\ 0 \\ 1 \\ r-3 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\begin{aligned} &\frac{(r-u)\hat{\tau}_1}{((u+v)\hat{\tau}_1+1)} \frac{((u+v+1)\hat{\tau}_2+1)(u+v+1)\hat{\tau}_2(r-v)\hat{\tau}_2}{((r-3)\hat{\tau}_2+r-1)((r-3)\hat{\tau}_2+r-2)((r-3)\hat{\tau}_2+r-3)} \times \\ &\qquad \frac{((r-3)\hat{\tau}_3+r-1)((r-3)\hat{\tau}_3+r-2)((r-3)\hat{\tau}_3+r-3)}{(u+v+2)\hat{\tau}_3(r-u)\hat{\tau}_3(r-v)\hat{\tau}_3} \\ = \; & 1\,. \end{aligned} \tag{4.71}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.71) gives

$$\frac{(u+v+1)\hat{\tau}_1}{((u+v)\hat{\tau}_1+1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Now let types $k$, $l$ and $m$ be

$$\begin{pmatrix} w \\ 0 \\ 0 \\ r-w \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix} \text{and} \begin{pmatrix} 2 \\ 0 \\ 1 \\ r-3 \end{pmatrix},$$

respectively, where $w \in \{1, \ldots, r-1\}$. Then (4.26) simplifies to

$$\frac{(r-u)\hat{\tau}_1}{((u+v+w)\hat{\tau}_1+1)} \frac{\Gamma((u+v+1)\hat{\tau}_2+2)(r-v)\hat{\tau}_2\Gamma((r-3)\hat{\tau}_2+r-3)}{\Gamma((u+v+1)\hat{\tau}_2+w)\Gamma((r-3)\hat{\tau}_2+r-w)} \times \frac{\Gamma((u+v+2)\hat{\tau}_3+w)\Gamma((r-3)\hat{\tau}_3+r-w)}{\Gamma((u+v+2)\hat{\tau}_3+1)(r-u)\hat{\tau}_3(r-v)\hat{\tau}_3\Gamma((r-3)\hat{\tau}_3+r-3)}$$
$$= 1. \tag{4.72}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.72) gives

$$\frac{(u+v+w+1)\hat{\tau}_1}{((u+v+w)\hat{\tau}_1+1)} = 1,$$

with solution

$$\hat{\tau}_1 = 1.$$

Assume $N = 4$. It is clear that the original histogram is of the form

$$\begin{bmatrix} r & 0 & 0 & 0 \\ 0 & r & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & 0 & 0 & r \end{bmatrix}.$$

Let the sub-histogram be

$$\begin{bmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \\ 0 & 0 & 0 \end{bmatrix}$$

and types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ r-3 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ 0 \\ 1 \\ r-3 \end{pmatrix},$$

respectively. Then (4.26) simplifies to

$$\frac{r\hat{\tau}_1}{(r\hat{\tau}_1+1)} \frac{((r+1)\hat{\tau}_2+1)(r+1)\hat{\tau}_2(r+1)\hat{\tau}_2}{((r-3)\hat{\tau}_2+r-1)((r-3)\hat{\tau}_2+r-2)((r-3)\hat{\tau}_2+r-3)} \times \frac{((r-3)\hat{\tau}_3+r-1)((r-3)\hat{\tau}_3+r-2)((r-3)\hat{\tau}_3+r-3)}{(r+2)\hat{\tau}_3 r\hat{\tau}_3(r+1)\hat{\tau}_3}$$
$$= 1. \tag{4.73}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.73) gives

$$\frac{(r+1)\hat{\tau}_1}{(r\hat{\tau}_1 + 1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Assume $K+1=5$. Then Equation (4.59) implies that $N \leq 2$. It is clear that the original histogram is some category permutation of the form

$$\begin{bmatrix} u_1 & 0 \\ u_2 & 0 \\ u_3 & 0 \\ 0 & v \\ 0 & r-v \end{bmatrix} \text{ or } \begin{bmatrix} u_1 & v_1 \\ u_2 & 0 \\ u_3 & 0 \\ 0 & v_2 \\ 0 & v_3 \end{bmatrix},$$

where $u_1, u_2, u_3, v_1, v_2, v_3 \in \{1, \ldots, r-2\}$, $v \in \{1, \ldots, r-1\}$, and $u_3 = r - u_1 - u_2$, $v_3 = r - v_1 - v_2$. Let the sub-histogram be

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 0 \\ 0 \end{bmatrix},$$

where $u_1, u_2, u_3 \in \{1, \ldots, r-2\}$ and $u_3 = r - u_1 - u_2$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ v \\ r-v \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ r-4 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \\ r-4 \end{pmatrix},$$

respectively, where $v \in \{1, \ldots, r-1\}$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{u_2\hat{\tau}_1}{(u_1\hat{\tau}_1+1)} \times \\ &\qquad \frac{\Gamma((u_1+1)\hat{\tau}_2+1)(u_1+1)\hat{\tau}_2(u_3+1)\hat{\tau}_2\Gamma(\hat{\tau}_2+1)\Gamma((r-4)\hat{\tau}_2+r-4)}{\Gamma(\hat{\tau}_2+v)\Gamma((r-4)\hat{\tau}_2+r-v)} \times \\ &\qquad \frac{\Gamma(\hat{\tau}_3+v)\Gamma((r-4)\hat{\tau}_3+r-v)}{(u_1+2)\hat{\tau}_3 u_2\hat{\tau}_3(u_3+1)\hat{\tau}_3\Gamma(\hat{\tau}_3+1)\Gamma((r-4)\hat{\tau}_3+r-4)} \\ = \; & 1\,. \end{aligned} \tag{4.74}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.74) gives

$$\frac{(u_1+1)\hat{\tau}_1}{(u_1\hat{\tau}_1+1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Now let types $k$, $l$ and $m$ be

$$\begin{pmatrix} v_1 \\ 0 \\ 0 \\ v_2 \\ v_3 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ r-4 \end{pmatrix} \text{and} \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \\ r-4 \end{pmatrix},$$

respectively, where $v_1, v_2, v_3 \in \{1, \ldots, r-2\}$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{u_2\hat{\tau}_1}{((u_1+v_1)\hat{\tau}_1+1)} \frac{\Gamma((u_1+1)\hat{\tau}_2+2)(u_3+1)\hat{\tau}_2\Gamma(\hat{\tau}_2+1)\Gamma((r-4)\hat{\tau}_2+r-4)}{\Gamma((u_1+1)\hat{\tau}_2+v_1)\Gamma(\hat{\tau}_2+v_2)\Gamma((r-4)\hat{\tau}_2+v_3)} \times \\ &\qquad \frac{\Gamma((u_1+2)\hat{\tau}_3+v_1)\Gamma(\hat{\tau}_3+v_2)\Gamma((r-4)\hat{\tau}_3+v_3)}{\Gamma((u_1+2)\hat{\tau}_3+1)u_2\hat{\tau}_3(u_3+1)\hat{\tau}_3\Gamma(\hat{\tau}_3+1)\Gamma((r-4)\hat{\tau}_3+r-4)} \\ = \; & 1\,. \end{aligned} \tag{4.75}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.75) gives

$$\frac{(u_1+1+v_1)\hat{\tau}_1}{((u_1+v_1)\hat{\tau}_1+1)} = 1\,,$$

with solution

$$\hat{\tau}_1 = 1\,.$$

Assume $K+1=6$. Then Equation (4.59) implies that $N \leq 2$. It is clear that the original histogram is some category permutation of the form

$$\begin{bmatrix} u_1 & 0 \\ u_2 & 0 \\ u_3 & 0 \\ 0 & v_1 \\ 0 & v_2 \\ 0 & v_3 \end{bmatrix},$$

where $u_1, u_2, u_3, v_1, v_2, v_3 \in \{1, \ldots, r-2\}$ and $u_3 = r - u_1 - u_2, v_3 = r - v_1 - v_2$. Let the sub-histogram be

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where $u_1, u_2, u_3 \in \{1, \ldots, r-2\}$ and $u_3 = r - u_1 - u_2$. Let types $k$, $l$ and $m$ be

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ r-5 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 1 \\ r-5 \end{pmatrix},$$

respectively, where $v_1, v_2, v_3 \in \{1, \ldots, r-2\}$. Then (4.26) simplifies to

$$\begin{aligned} &\frac{u_2\hat{\tau}_1}{(u_1\hat{\tau}_1 + 1)} \times \\ &\frac{((u_1+1)\hat{\tau}_2 + 1)(u_1+1)\hat{\tau}_2(u_3+1)\hat{\tau}_2\Gamma(\hat{\tau}_2+1)\Gamma(\hat{\tau}_2+1)\Gamma((r-5)\hat{\tau}_2 + r - 5)}{\Gamma(\hat{\tau}_2 + v_1)\Gamma(\hat{\tau}_2 + v_2)\Gamma((r-5)\hat{\tau}_2 + v_3)} \times \\ &\frac{\Gamma(\hat{\tau}_3 + v_1)\Gamma(\hat{\tau}_3 + v_2)\Gamma((r-5)\hat{\tau}_3 + v_3)}{(u_1+2)\hat{\tau}_3 u_2 \hat{\tau}_3 (u_3+1)\hat{\tau}_3\Gamma(\hat{\tau}_3+1)\Gamma(\hat{\tau}_3+1)\Gamma((r-5)\hat{\tau}_3 + r - 5)} \\ = \; & 1 . \end{aligned} \tag{4.76}$$

We have already shown that $\hat{\tau}_2 = \hat{\tau}_3 = 1$. Substituting these values into (4.76) gives

$$\frac{(u_1+1)\hat{\tau}_1}{(u_1\hat{\tau}_1 + 1)} = 1 ,$$

with solution

$$\hat{\tau}_1 = 1 .$$

Assume $K + 1 \geq 7$. Then Equation (4.59) implies that $N \leq 1$.

Hence all of the special cases have been dealt with.

Finally, in light of Theorem 4.3.4, Lemma 4.3.3 and the fact that your predictive probabilities are strictly positive then allow Theorem 2.4.1 to complete the proof.□

**Example 4.3.2** *To illustrate how Theorem 4.3.6 works, consider the case* $N+1 = 4$, $r = 5$, $K+1 = 4$ *and suppose that it is desired to prove that the* $\hat{\tau}$ *variable associated with the histogram*

$$\begin{bmatrix} 4 & 1 & 0 & 2 \\ 0 & 1 & 4 & 3 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix} \tag{4.77}$$

*must equal* 15. *Firstly, let* $\hat{\tau}_1, \hat{\tau}_2$ *and* $\hat{\tau}_3$ *represent the* $\hat{\tau}$ *variables associated with the histograms*

$$\begin{bmatrix} 4 & 1 & 0 & 1 \\ 0 & 1 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 3 \end{bmatrix}, \begin{bmatrix} 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 3 \end{bmatrix} \textit{ and } \begin{bmatrix} 4 & 1 & 0 & 2 \\ 0 & 1 & 4 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 3 \end{bmatrix},$$

*respectively, each scaled by a factor of* $1/15$. *Note that* $\hat{\tau}_2 = \hat{\tau}_1$. *The 3-cycle formed by these histograms will be satistied if and only if*

$$\begin{aligned} \frac{7(5\hat{\tau}_1 + 1)}{6(6\hat{\tau}_1 + 1)} &= 1 \\ \Rightarrow \hat{\tau}_1 &= 1 \end{aligned}$$

*(see Equation (4.27)). Now let* $\hat{\tau}_1, \hat{\tau}_2$ *and* $\hat{\tau}_3$ *represent the* $\hat{\tau}$ *variables associated with the histograms*

$$\begin{bmatrix} 4 & 1 & 0 & 3 \\ 0 & 1 & 4 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 4 & 1 & 0 & 3 \\ 0 & 1 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 \end{bmatrix} \textit{ and } \begin{bmatrix} 4 & 1 & 0 & 3 \\ 0 & 1 & 4 & 2 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix}, \tag{4.78}$$

*respectively, each scaled by a factor of* $1/15$. *The 3-cycle formed by these histograms will be satistied if and only if*

$$\begin{aligned} \frac{7\hat{\tau}_1(5\hat{\tau}_2 + 1)}{6(6\hat{\tau}_1 + 1)\hat{\tau}_2} &= 1 \\ \Rightarrow \hat{\tau}_1 &= \frac{6\hat{\tau}_2}{-\hat{\tau}_2 + 7} \end{aligned} \tag{4.79}$$

*(see Equations (4.27) and (4.28)). Leaving* $\hat{\tau}_1$ *and* $\hat{\tau}_2$ *unchanged, now let* $\hat{\tau}_3$ *be the*

*$\hat{\tau}$ variable associated with the histogram*

$$\begin{bmatrix} 4 & 1 & 0 & 3 \\ 0 & 1 & 4 & 0 \\ 0 & 1 & 1 & 2 \\ 1 & 2 & 0 & 0 \end{bmatrix},$$

*again scaled by a factor of* $1/15$. *The 3-cycle formed by the first two histograms given in (4.78) and this one will be satisfied if and only if*

$$\begin{aligned} \frac{3\hat{\tau}_1(3\hat{\tau}_2+1)}{4(2\hat{\tau}_1+1)\hat{\tau}_2} &= 1 \\ \Rightarrow \hat{\tau}_1 &= \frac{4\hat{\tau}_2}{\hat{\tau}_2+3} \end{aligned} \tag{4.80}$$

*(see Equations (4.29) and (4.30)). Together (4.79) and (4.80) give a quadratic equation in* $\hat{\tau}_2$ *with roots* $0$, $1$. *Substituting the only positive solution,* $\hat{\tau}_2 = 1$, *into either of these equations gives* $\hat{\tau}_1 = 1$. *Finally, let* $\hat{\tau}_1, \hat{\tau}_2$ *and* $\hat{\tau}_3$ *represent the* $\hat{\tau}$ *variables associated with the histograms*

$$\begin{bmatrix} 4 & 1 & 0 & 1 \\ 0 & 1 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 3 \end{bmatrix}, \begin{bmatrix} 4 & 1 & 0 & 3 \\ 0 & 1 & 4 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 4 & 1 & 0 & 2 \\ 0 & 1 & 4 & 3 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix},$$

*respectively, each scaled by a factor of* $1/15$. *We have just shown that* $\hat{\tau}_1 = \hat{\tau}_2 = 1$. *Hence, the 3-cycle formed by these histograms will be satisfied if and only if*

$$\begin{aligned} \frac{18(3\hat{\tau}_3+2)(3\hat{\tau}_3+1)}{5(7\hat{\tau}_3+2)(7\hat{\tau}_3+1)} &= 1 \\ \Rightarrow \hat{\tau}_3 &= 1, \frac{-26}{83} \end{aligned}$$

*(see Equation (4.46)). Clearly, only the first of these roots is positive and the (unscaled)* $\hat{\tau}$ *variable associated with the histogram in (4.77) must equal* $15$.

## 4.4 Coherency of Mosimann, Brier, *et al.*

The results of the preceding section shed some light on the reasons for the coherency requirement to give zero probability to using the strategy given by (4.16)

(Mosimann) or (4.19) (Brier) for certain values of $N+1$, $r$ and $K+1$. It is clear that $\mathcal{H}_M \subseteq \mathcal{H}_P$ and $\mathcal{H}_B \subseteq \mathcal{H}_P$. Suppose $h, \mathfrak{h} \in \mathcal{H}_P$ are assigned the same $\hat{\tau}$ variable, *i.e.*, $\mathfrak{h}$ can be obtained from $h$ by permuting rows (and columns, if necessary). Obviously, Brier's $\hat{C}$, and therefore $\hat{\tau}$, for $h$ and $\mathfrak{h}$ are identical (see Equation (4.17)). This is also true of Mosimann's $\hat{C}$, and therefore $\hat{\tau}$, as we now show.

**Theorem 4.4.1** *Let $h, \mathfrak{h} \in \mathcal{H}_P$ be positive histograms such that one can be obtained from the other by permuting rows (and columns, if necessary). Then Mosimann's $\hat{C}$ for $h$ and $\mathfrak{h}$ are identical.*

PROOF: Specifically, suppose that the $j^{\text{th}}$ row of $h$ is the $\pi(j)^{\text{th}}$ row of $\mathfrak{h}$, $j = 1, \ldots, K+1$, for some permutation, $\pi$, of the numbers $\{1, \ldots, K+1\}$. Consider $W_{\text{Mult}} = [w_{jk}]$ of $h$, formed using data from the first $K$ categories of $h$ only. Then $w_{jk}$ is a function of the data from categories $j$ and $k$ of $h$, $j, k \in \{1, \ldots, K\}$. Consider $W'_{\text{Mult}} = [w'_{jk}]$ of $\mathfrak{h}$, formed using data from categories $\pi(1), \ldots, \pi(K)$ of $\mathfrak{h}$. Then $w'_{jk}$ is a function of the data from categories $j$ and $k$ of $\mathfrak{h}$, or categories $\pi^{-1}(j), \pi^{-1}(k)$ of $h$, $\pi^{-1}(j), \pi^{-1}(k) \in \{1, \ldots, K\}$. Let $R$ be the permutation matrix that results from applying $\pi$ to the rows of the $(K+1)$-dimensional identity matrix. Let $P$ be the matrix obtained by removing row $\pi(K+1)$ and the last column from $R$. Consider the matrix $Q = PW_{\text{Mult}}P^T$. Then

$$\begin{aligned} Q_{jk} &= (PW_{\text{Mult}}P^T)_{jk} \\ &= (PW_{\text{Mult}})_{j\pi^{-1}(k)} \\ &= w_{\pi^{-1}(j)\pi^{-1}(k)} \\ &= w'_{jk}. \end{aligned}$$

Hence, $Q = W'_{\text{Mult}}$ and

$$\begin{aligned} |W'_{\text{Mult}}| &= |PW_{\text{Mult}}P^T| \\ &= |W_{\text{Mult}}|. \end{aligned}$$

Similarly, $|W'_{\text{DMD}}| = |W_{\text{DMD}}|$, where the former matrix pertains to $\mathfrak{h}$ and the latter to $h$. □

Hence, both strategies (4.16) and (4.19) may be thought of as subcases of (4.3) using (4.22). This suggests that it is precisely because neither Brier nor Mosimann's

$\hat{\tau}$ generally equals the value $rN$ that their estimators produce effectively incoherent predictive probabilities in some cases. Care must be taken in extending the results of the previous section, however, as the sets $\mathcal{H}_M$ and $\mathcal{H}_B$ are not guaranteed to be linked and neither is the existence of an unsatisfied $n$-cycle in all linked subsets guaranteed. Particularly for large $N$, it seems likely from inspection of the formulae for Mosimann and Brier's $\hat{C}$ that neighbouring histograms, that is, histograms that differ in exactly one $\underline{Y}^{(i)}$, would either both be members of $\mathcal{H}_M$ or $\mathcal{H}_B$, or both not be. This may lead to overall linking of these sets. Also, it appears to be relatively easy to construct $n$-cycles that require certain $\hat{\tau}$ to equal $rN$ in order to be satisfied. It is therefore conjectured, for $r \geq K+1$ and $N+1$ sufficiently large, that the only way for you to be coherent in using either the strategy given by (4.16) or (4.19) is to give zero probability to exactly those situations in which it would apply.

Although no coherency analysis was directly undertaken involving pseudo-maximum likelihood estimates of $\underline{\alpha}$ or any of the other estimators of $C$ or $\tau$ mentioned at the end of §4.1, the fact that they, too, all employ (4.22) may lead to their downfall if considered via (4.3) as a strategy for estimating predictive probabilities.

## 4.5 Alternative Estimates of $\underline{\alpha}$

Up to this point we have concentrated on estimators of $\underline{\alpha}$ that employ (4.22). Obviously there are many other possible estimators that do not share this property, and a few of these will now be discussed.

### 4.5.1 Generalising the Frequency Mimicking Partition of $\hat{\tau}$

A natural generalisation of the strategy in (4.22) would be to assert

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{N} Y_j^{(i)} + c}{rN + (K+1)c}\hat{\tau}, \qquad j = 1, \ldots, K+1, \tag{4.81}$$

where $c > 0, c \in \mathbb{R}$. Similarly, another possibility would be to assert

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{N} Y_j^{(i)} + c_j}{rN + \sum_{k=1}^{K+1} c_k}\hat{\tau}, \qquad j = 1, \ldots, K+1, \tag{4.82}$$

where $c_j > 0, c_j \in \mathbb{R}, j = 1, \ldots, K+1$.

Again the astute reader may have noticed a similarity between the expressions in (4.81) and (4.82) and a posterior expectation associated with a Dirichlet distribution (symmetric in the case of (4.81)). This relationship will be explored in Chapter 5.

Empirical studies along the lines of those previously conducted assuming (4.22) suggest that if you want to be coherent in giving a nonzero probability to using a strategy that estimates your predictive probabilities using (4.3) with (4.81), then you are, in fact, very restricted in the way you estimate $\tau$. Specifically, for given values of $N+1$, $r$ and $K+1$ it seems that your estimates of $\tau$ corresponding to most sorts of positive histograms must all equal $rN+(K+1)c$. Furthermore, if $r \geq K+1$ this may be true for *all* positive histograms. Note that (4.81) and $\hat{\tau} = rN+(K+1)c$ together imply

$$\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)} + c, \qquad j = 1, \ldots, K+1\,. \tag{4.83}$$

**Theorem 4.5.1** *If you assert your predictive probabilities using (4.3) where* $\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)} + c$, $j = 1, \ldots, K+1$, *for all* $H \in \mathcal{H}_P$, *the set of positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Follow the proof of Theorem 4.3.4 but replace $S_{j,h}$ by $S_{j,h} + c$ and $S_{j,\mathfrak{h}}$ by $S_{j,\mathfrak{h}} + c$ in all of the equations. □

**Theorem 4.5.2** *Let* $N+1 \geq 3, r \geq K+1$ *and* $K+1 \geq 3$. *If you assert your predictive probabilities using (4.3) where*

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{N} Y_j^{(i)} + c}{rN + (K+1)c}\,\hat{\tau}, \qquad j = 1, \ldots, K+1\,,$$

*for all* $H \in \mathcal{H}_P$, *the set of positive histograms, then all of the* $\hat{\tau}$ *variables must equal* $rN + (K+1)c$ *to allow a nonzero solution for the components of* $\underline{q}$ *that appear in the system of coherency induced equations.*

PROOF: Appropriate adjustments to the proof of Theorem 4.3.6 give the required result. □

Also, empirical studies along the lines of those previously conducted assuming (4.22) suggest that if you want to be coherent in giving a nonzero probability to using a strategy that estimates your predictive probabilities using (4.3) with (4.82),

then you are, in fact, very restricted in the way you estimate $\tau$. Specifically, for given values of $N+1$, $r$ and $K+1$ it seems that your estimates of $\tau$ corresponding to most sorts of positive histograms must all equal $rN + \sum_{k=1}^{K+1} c_k$. Furthermore, if $r \geq K+1$ this may be true for *all* positive histograms. Note that (4.82) and $\hat{\tau} = rN + \sum_{k=1}^{K+1} c_k$ together imply

$$\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)} + c_j, \qquad j = 1, \ldots, K+1\,. \tag{4.84}$$

**Theorem 4.5.3** *If you assert your predictive probabilities using (4.3) where* $\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)} + c_j$, $j = 1, \ldots, K+1$, *for all* $H \in \mathcal{H}_P$, *the set of positive histograms, then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Follow the proof of Theorem 4.3.4 but replace $S_{j,h}$ by $S_{j,h} + c_j$ and $S_{j,\mathfrak{h}}$ by $S_{j,\mathfrak{h}} + c_j$ in all of the equations. □

**Theorem 4.5.4** *Let* $N+1 \geq 3, r \geq K+1$ *and* $K+1 \geq 3$. *If you assert your predictive probabilities using (4.3) where*

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{N} Y_j^{(i)} + c_j}{rN + \sum_{k=1}^{K+1} c_k}\,\hat{\tau}, \qquad j = 1, \ldots, K+1\,,$$

*for all* $H \in \mathcal{H}_P$, *the set of positive histograms, then all of the* $\hat{\tau}$ *variables must equal* $rN + \sum_{k=1}^{K+1} c_k$ *to allow a nonzero solution for the components of* $\underline{q}$ *that appear in the system of coherency induced equations.*

PROOF: Appropriate adjustments to the proof of Theorem 4.3.6 give the required result. □

Note that (4.81) and (4.82) will always produce valid estimates of $\underline{\alpha}$, being strictly positive, so that Theorems 4.5.2 and 4.5.4 hold equally well if they refer to a strategy that asserts these estimators for *all* possible conditioning histograms, given that this set of histograms is known to be linked.

Also, as the constant, $c$, in (4.81) increases, the estimates tend towards

$$\hat{\alpha}_j = \frac{\tau}{K+1}, \qquad j = 1, \ldots, K+1\,,$$

in the extreme ($c \to \infty$).

### 4.5.2 Maximum Likelihood Estimates

This approach to parameter estimation relies on the fact that, in the presence of observed data, the probability distribution of the data may be viewed as a likelihood function for the prior. Maximising this likelihood with respect to the parameters of the prior then amounts to selecting the most 'plausible' prior in the light of the data at hand. Maximum likelihood estimation (MLE) of the parameter $\underline{\alpha}$ in the Dirichlet-Multinomial distribution leads naturally to estimation of the mean parameters $\lambda_j \equiv \alpha_j/\tau$, $j = 1, \ldots, K$, and the scale parameter $\tau$. Consider

$$\begin{aligned} L(\underline{\lambda}, \tau) &\equiv \prod_{i=1}^{N} L\left(\underline{\lambda}, \tau \,\middle|\, \underline{Y}^{(i)}\right) \\ &= \prod_{i=1}^{N} \left[ \frac{r!\Gamma(\tau) \prod_{j=1}^{K+1} \Gamma\left(\tau\lambda_j + Y_j^{(i)}\right)}{\prod_{j=1}^{K+1} \left[Y_j^{(i)}!\right] \prod_{j=1}^{K+1} [\Gamma(\tau\lambda_j)] \Gamma(\tau + r)} \right], \end{aligned}$$

where $\lambda_{K+1} = 1 - \sum_{j=1}^{K} \lambda_j$. For $r > 1$, differentiating the logarithm of this expression with respect to $\lambda_j$, $j = 1, \ldots, K$, and $\tau$, and setting equal to zero gives

$$\begin{aligned} \sum_{i=1}^{N} \sum_{l=0}^{Y_j^{(i)}-1} \frac{1}{l + \tau\lambda_j} &= \sum_{i=1}^{N} \sum_{l=0}^{Y_{K+1}^{(i)}-1} \frac{1}{l + \tau\lambda_{K+1}}, \qquad j = 1, \ldots, K, \\ \sum_{i=1}^{N} \sum_{j=1}^{K+1} \sum_{l=0}^{Y_j^{(i)}-1} \frac{\lambda_j}{l + \tau\lambda_j} &= N \sum_{l=0}^{r-1} \frac{1}{l + \tau} \end{aligned} \tag{4.85}$$

(see Danaher [21], p1779 where there are some misprints). When $r = 1$, the likelihood equations for $\lambda_j$ reduce to

$$\lambda_j = \frac{x_j}{x_{K+1}} \lambda_{K+1}, \qquad j = 1, \ldots, K,$$

where $x_j = \sum_{i=1}^{N} \left(Y_j^{(i)} = 1\right)$ is the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $j$. However $\tau$ does not feature at all. This reinforces the idea that $r > 1$ is needed for identifiability of $\tau$.

**Example 4.5.1** *To illustrate the calculation of $\hat{\alpha}$ by maximum likelihood, consider again the data from Mosimann [72], reproduced in Table B.1 of Appendix B, §B.1. For this data, recall $N = 73$, $r = 100$, $K + 1 = 4$. Solving*

$$\sum_{i=1}^{N} \sum_{l=0}^{Y_j^{(i)}-1} \frac{1}{l + \tau\lambda_j} = \sum_{i=1}^{N} \sum_{l=0}^{Y_4^{(i)}-1} \frac{1}{l + \tau\lambda_4}, \qquad j = 1, \ldots, 3,$$

$$\lambda_4 = 1 - \lambda_1 - \lambda_2 - \lambda_3$$

$$\sum_{i=1}^{N}\sum_{j=1}^{4}\sum_{l=0}^{Y_j^{(i)}-1}\frac{\hat{\lambda}_j}{l+\tau\hat{\lambda}_j} = 73\sum_{l=0}^{99}\frac{1}{l+\tau}$$

*gives*

$$\begin{aligned}\hat{\lambda}_1 &= 0.862 \quad (3\ d.p.)\\ \hat{\lambda}_2 &= 0.016 \quad (3\ d.p.)\\ \hat{\lambda}_3 &= 0.089 \quad (3\ d.p.)\\ \hat{\lambda}_4 &= 0.032 \quad (3\ d.p.)\\ \hat{\tau} &= 60.195 \quad (3\ d.p.),\end{aligned}$$

*or*

$$\begin{aligned}\hat{\alpha}_1 &= 51.895 \quad (3\ d.p.)\\ \hat{\alpha}_2 &= 0.989 \quad (3\ d.p.)\\ \hat{\alpha}_3 &= 5.345 \quad (3\ d.p.)\\ \hat{\alpha}_4 &= 1.966 \quad (3\ d.p.).\end{aligned}$$

For the special case of the Beta-Binomial distribution ($K = 1$), some use has been made of maximum likelihood (Griffiths [46], Williams [90], Segreti and Munson [82]), however, when $K > 1$ obtaining a solution to (4.85) can be computationally unattractive. For this reason, pseudo maximum likelihood estimation (see § 4.2) may be preferred.

Due to the numerical difficulties of obtaining a solution to the likelihood equations (4.85) and the inexact nature of such a solution, it is not feasible to carry out a study of the coherency implications of a strategy that would use (4.3) with $\underline{\hat{\alpha}}$ obtained via MLE.

### 4.5.3 The Marginal Approach

Another approach that uses the marginal distribution of $\underline{Y}^{(i)}$ would be to equate the observed proportions of the $N$ groups resulting in each possible type outcome with the theoretical probabilities involving the parameter $\underline{\alpha}$ and solve the resulting

system of equations. For example,

$$\begin{aligned} P\left(\underline{Y}^{(i)} = \text{type } 1 \,\middle|\, \underline{\alpha}\right) &= \frac{r!\Gamma(\tau)\Gamma(\alpha_1 + r)\prod_{j=2}^{K+1}\Gamma(\alpha_j)}{r!0!^K \prod_{j=1}^{K+1}[\Gamma(\alpha_j)]\Gamma(\tau + r)} \\ &= \frac{\Gamma(\tau)\Gamma(\alpha_1 + r)}{\Gamma(\alpha_1)\Gamma(\tau + r)} \\ &= \frac{(\alpha_1 + r - 1)\cdots\alpha_1}{(\tau + r - 1)\cdots\tau}, \end{aligned}$$

so that the first equation would be

$$\frac{(\alpha_1 + r - 1)\cdots\alpha_1}{(\tau + r - 1)\cdots\tau} = \frac{x_1}{N},$$

where $x_1$ is the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type 1. However, unless $r = 1$, this system of $\left({}^{r+K}C_K - 1\right)$ linearly independent equations in $K + 1$ unknowns will generally be overdetermined and have no solution. One possibility would be to minimise the sum of the squared differences in the left and right-hand sides of the equations, with respect to $\underline{\alpha}$. Again, the details of the coherency implications of a strategy that would use (4.3) with $\hat{\underline{\alpha}}$ estimated in this way are considered too complex to pursue.

### 4.5.4 Symmetric Dirichlet Distributions

Many authors, including Lindley [70], Good [38] and Block and Watson [9], have considered the use of symmetric Dirichlet priors for describing vectors of probabilities. These are Dirichlet($\underline{\alpha}$) distributions where $\alpha_1 = \cdots = \alpha_{K+1} = k$, say. The problem of estimating $\underline{\alpha}$ in this case reduces from one involving $K + 1$ parameters to the estimation of a single value. In this situation, MLE would seem the most likely choice of approach. Note then that Theorem 4.2.1 with $\lambda_1 = \cdots = \lambda_{K+1} = 1/(K+1)$ gives a sufficient condition for the existence of a finite MLE of $k$. Alternatively, perhaps some estimate of the 'repeat rate'

$$\left(\theta_1^{(i)}\right)^2 + \left(\theta_2^{(i)}\right)^2 + \cdots + \left(\theta_{K+1}^{(i)}\right)^2$$

may be obtained from the data and equated to its expectation

$$\frac{k+1}{(K+1)k+1}.$$

It may be possible to carry out the sort of analysis described in §4.3 for the case of a symmetric Dirichlet distribution and thus discover exactly what strategies for estimating the parameter $\underline{\alpha} = k\underline{1}$ lead to coherent predictive probabilities using (4.3).

Having completed a thorough study of the ways in which an empirical Bayes statistician may deal with the unknown parameter $\underline{\alpha}$, we will now turn our attention to the hierarchical Bayesian approach.

# Chapter 5

# Hierarchical Bayesian Estimates

In this chapter the coherency properties of hierarchical Bayesian estimates of predictive probabilities will be investigated. An alternative formulation of the problem is also considered that leads to a possible explanation for the coherent probability specification strategies discovered in Chapters 3 and 4.

The hierarchical Bayesian approach to the estimation of predictive probabilities given is based upon the *complete* specification of probability distributions that describe the quantities involved in the problem. We will continue to assume that

$$\begin{aligned} f\left(\underline{Y}^{(i)} \middle| \underline{\theta}^{(i)}\right) &= \frac{r!}{\prod_{j=1}^{K+1} Y_j^{(i)}!} \prod_{j=1}^{K+1} \left[\theta_j^{(i)}\right]^{Y_j^{(i)}} \\ &\sim \text{Multinomial}\left(r, \underline{\theta}^{(i)}\right), \qquad i = 1, \ldots, N+1, \end{aligned}$$

and

$$\begin{aligned} f\left(\underline{\theta}^{(i)} \middle| \underline{\alpha}\right) &= \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \prod_{j=1}^{K+1} \left[\theta_j^{(i)}\right]^{\alpha_j - 1} \\ &\sim \text{Dirichlet}(\underline{\alpha}), \qquad i = 1, \ldots, N+1, \end{aligned}$$

so that

$$\begin{aligned} f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right) &= \frac{r!\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right) \prod_{j=1}^{K+1} \Gamma\left(\alpha_j + Y_j^{(i)}\right)}{\prod_{j=1}^{K+1} \left[Y_j^{(i)}!\right] \prod_{j=1}^{K+1} [\Gamma(\alpha_j)] \Gamma\left(\sum_{j=1}^{K+1} \alpha_j + r\right)} \\ &\sim \text{DMD}(r, \underline{\alpha}), \qquad i = 1, \ldots, N+1, \end{aligned}$$

where $\underline{\alpha} > \underline{0}$. The $\underline{Y}^{(i)}$ vectors, $i = 1, \ldots, N+1$, are presumed independent conditional upon $\underline{\alpha}$, a structure the subjectivist would recognise as motivating an exchangeable distribution over the $\underline{Y}^{(i)}$. It now remains to describe the hyperparameter, $\underline{\alpha}$, with a probability distribution, $f(\underline{\alpha})$. Then, the predictive distribution for $\underline{Y}^{(N+1)}$ is given by

$$f\left(\underline{Y}^{(N+1)} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) = \int_{\underline{\alpha}} f\left(\underline{Y}^{(N+1)} \middle| \underline{\alpha}\right) f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) d\underline{\alpha},$$

where

$$\begin{aligned} f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) &= \frac{f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} \middle| \underline{\alpha}\right) f(\underline{\alpha})}{f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)} \\ &= \frac{\prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha})}{\int_{\underline{\alpha}} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha}) d\underline{\alpha}} \end{aligned}$$

is the posterior density for $\underline{\alpha}$ given the data. That is,

$$f\left(\underline{Y}^{(N+1)} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) = \int_{\underline{\alpha}} \frac{\prod_{i=1}^{N+1}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha})}{\int_{\underline{\alpha}} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha}) d\underline{\alpha}} d\underline{\alpha},$$

or

$$\begin{aligned} & P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) \\ = & \int_{\underline{\alpha}} \frac{P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{\alpha}\right) \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha})}{\int_{\underline{\alpha}} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha}) d\underline{\alpha}} d\underline{\alpha} \qquad (5.1) \\ = & \int_{\underline{\alpha}} \frac{\frac{r!\Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j\right) \prod_{j=1}^{K+1} \Gamma(\hat{\alpha}_j + y_j)}{\prod_{j=1}^{K+1}[y_j!] \prod_{j=1}^{K+1}[\Gamma(\hat{\alpha}_j)]\Gamma\left(\sum_{j=1}^{K+1} \hat{\alpha}_j + r\right)} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha})}{\int_{\underline{\alpha}} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)\right] f(\underline{\alpha}) d\underline{\alpha}} d\underline{\alpha}. \end{aligned}$$

It remains to study some of the possible choices for $f(\underline{\alpha})$ and their effects on the coherency of the resulting predictive probabilities.

## 5.1 Noninformative Priors

In some situations there may be little or no prior information available concerning $\underline{\alpha}$. What is needed then is a prior which 'favours' no possible values of $\underline{\alpha}$ over any others. It frequently happens that a noninformative prior is an improper prior, namely one

which has infinite mass, but this does not necessarily preclude the posterior from being a proper distribution.

**Notation**: Throughout the remainder of this chapter, the expression 'const.' refers to a strictly positive, real-valued constant.

### 5.1.1 Hyperprior Identically Equal to 1: $f(\underline{\alpha}) \equiv 1$

A commonly employed noninformative prior, used by Laplace [62], is to equate the hyperprior density identically equal to unity. However, it will be shown that this improper prior leads to an undefined posterior for $\underline{\alpha}$ in the context of the problem here. We first prove the following lemma.

**Lemma 5.1.1** *If* $d \geq 1$, $J \geq 2$, $d, J \in \mathbb{N}$ *and* $s_j \geq 0$, $s_j \in \mathbb{N}$, $j = 1, \ldots, J$, *then*

$$\int_0^\infty \cdots \int_0^\infty \frac{\prod_{j=1}^J x_j^{s_j}}{\left(\sum_{j=1}^J x_j + d\right)^{\left(\sum_{j=1}^J s_j + J\right)}} dx_J \cdots dx_1$$

$$\geq \text{ const.} \int_0^\infty \cdots \int_0^\infty \frac{\prod_{j=1}^{J-1} x_j^{s_j}}{\left(\sum_{j=1}^{J-1} x_j + d\right)^{\left(\sum_{j=1}^{J-1} s_j + J - 1\right)}} dx_{J-1} \cdots dx_1 \,.$$

PROOF:

$$\begin{aligned}
&\text{LHS} \\
&= \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{J-1} \left[x_j^{s_j}\right] \int_0^\infty \frac{x_J^{s_J}}{\left(\sum_{j=1}^J x_j + d\right)^{\left(\sum_{j=1}^J s_j + J\right)}} dx_J dx_{J-1} \cdots dx_1 \\
&\geq \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{J-1} \left[x_j^{s_j}\right] \int_{\sum_{j=1}^{J-1} x_j + d}^\infty \frac{x_J^{s_J}}{\left(\sum_{j=1}^J x_j + d\right)^{\left(\sum_{j=1}^J s_j + J\right)}} dx_J dx_{J-1} \cdots dx_1 \\
&\geq \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{J-1} \left[x_j^{s_j}\right] \int_{\sum_{j=1}^{J-1} x_j + d}^\infty \frac{x_J^{s_J}}{(2x_J)^{\left(\sum_{j=1}^J s_j + J\right)}} dx_J dx_{J-1} \cdots dx_1 \\
&= \text{const.} \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{J-1} x_j^{s_j} \left[ \frac{-1}{x_J^{\left(\sum_{j=1}^{J-1} s_j + J - 1\right)}} \right]_{\sum_{j=1}^{J-1} x_j + d}^\infty dx_{J-1} \cdots dx_1 \\
&= \text{const.} \int_0^\infty \cdots \int_0^\infty \frac{\prod_{j=1}^{J-1} x_j^{s_j}}{\left(\sum_{j=1}^{J-1} x_j + d\right)^{\left(\sum_{j=1}^{J-1} s_j + J - 1\right)}} dx_{J-1} \cdots dx_1 \\
&= \text{RHS}
\end{aligned}$$

□

**Theorem 5.1.2** *If*

$$f\left(\underline{Y}^{(i)}\middle|\underline{\alpha}\right) \sim DMD(r,\underline{\alpha}), \qquad i=1,\ldots,N+1,$$

*and*

$$f(\underline{\alpha}) \equiv 1,$$

*then* $f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ *is infinite and hence the posterior distribution for* $\underline{\alpha}$, $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$, *is undefined, for every* $\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}$.

PROOF: For $j=1,\ldots,K+1$, let $S_j=\sum_{i=1}^{N} Y_j^{(i)}$. Then recursive use of Lemma 5.1.1 shows that

$$\begin{aligned}
& f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right) \\
= & \int_{\underline{\alpha}} \prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)}\middle|\underline{\alpha}\right)\right] f(\underline{\alpha})d\underline{\alpha} \\
= & \int_{\underline{\alpha}} \prod_{i=1}^{N}\left[\frac{r!\Gamma\left(\sum_{j=1}^{K+1}\alpha_j\right)\prod_{j=1}^{K+1}\Gamma\left(\alpha_j+Y_j^{(i)}\right)}{\prod_{j=1}^{K+1}\left[Y_j^{(i)}!\right]\prod_{j=1}^{K+1}[\Gamma(\alpha_j)]\Gamma\left(\sum_{j=1}^{K+1}\alpha_j+r\right)}\right] 1\, d\underline{\alpha} \\
= & \text{ const.} \int_{\underline{\alpha}} \prod_{i=1}^{N}\left[\frac{\prod_{j=1}^{K+1}\Gamma\left(\alpha_j+Y_j^{(i)}\right)}{\prod_{j=1}^{K+1}[\Gamma(\alpha_j)]\left(\sum_{j=1}^{K+1}\alpha_j+r-1\right)\cdots\left(\sum_{j=1}^{K+1}\alpha_j\right)}\right] d\underline{\alpha} \\
\geq & \text{ const.} \int_{\underline{\alpha}} \frac{\prod_{j=1}^{K+1}\alpha_j^{S_j}}{\left(\sum_{j=1}^{K+1}\alpha_j+r-1\right)^{rN}} d\underline{\alpha} \\
\geq & \text{ const.} \int_{\underline{\alpha}} \frac{\prod_{j=1}^{K+1}\alpha_j^{S_j}}{\left(\sum_{j=1}^{K+1}\alpha_j+r-1\right)^{(rN+K+1)}} d\underline{\alpha} \\
\geq & \text{ const.} \int_0^\infty \cdots \int_0^\infty \frac{\prod_{j=1}^{K}\alpha_j^{S_j}}{\left(\sum_{j=1}^{K}\alpha_j+r-1\right)^{\left(\sum_{j=1}^{K}S_j+K\right)}} d\alpha_K\cdots d\alpha_1 \\
& \vdots \\
\geq & \text{ const.} \int_0^\infty \frac{\alpha_1^{S_1}}{(\alpha_1+r-1)^{(S_1+1)}} d\alpha_1 \\
\geq & \text{ const.} \int_{r-1}^\infty \frac{\alpha_1^{S_1}}{(\alpha_1+r-1)^{(S_1+1)}} d\alpha_1 \\
\geq & \text{ const.} \int_{r-1}^\infty \frac{\alpha_1^{S_1}}{(2\alpha_1)^{(S_1+1)}} d\alpha_1 \\
= & \text{ const.} \int_{r-1}^\infty \frac{1}{\alpha_1} d\alpha_1 \\
= & \infty .
\end{aligned}$$

□

The lack of invariance of the constant prior under reparameterisations of a problem has come under severe criticism. This led to a search, pioneered by Jeffreys, for noninformative priors which *are* appropriately invariant under transformations.

### 5.1.2 Jeffreys' Hyperprior

A widely used method for determining noninformative priors in a general setting is that of Jeffreys [52]. He suggests the use of

$$f(\underline{\alpha}) = (|I(\underline{\alpha})|)^{1/2},$$

where $I(\underline{\alpha})$ is the $(K+1) \times (K+1)$ Fisher information matrix. Under commonly satisfied assumptions, this is the matrix with entry

$$[I(\underline{\alpha})]_{kl} = -E\left[\frac{\partial^2}{\partial\alpha_k \partial\alpha_l} \ln f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} | \underline{\alpha}\right)\right]$$

in row $k$, column $l$. Now,

$$f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} \,\middle|\, \underline{\alpha}\right) = \prod_{i=1}^{N} f\left(\underline{Y}^{(i)} \,\middle|\, \underline{\alpha}\right),$$

so that

$$\ln f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} \,\middle|\, \underline{\alpha}\right) = \sum_{i=1}^{N}\left[\ln r! - \sum_{j=1}^{K+1} \ln Y_j^{(i)}! + \right.$$

$$\left. \ln\Gamma\left(\sum_{j=1}^{K+1}\alpha_j\right) - \ln\Gamma\left(\sum_{j=1}^{K+1}\alpha_j + r\right) + \sum_{j=1}^{K+1}\ln\Gamma\left(\alpha_j + Y_j^{(i)}\right) - \sum_{j=1}^{K+1}\ln\Gamma(\alpha_j)\right].$$

Then,

$$\begin{aligned} &\frac{\partial}{\partial\alpha_k} \ln f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} \,\middle|\, \underline{\alpha}\right) \\ &= N\left(\psi\left(\sum_{j=1}^{K+1}\alpha_j\right) - \psi\left(\sum_{j=1}^{K+1}\alpha_j + r\right) - \psi(\alpha_k)\right) + \sum_{i=1}^{N}\psi\left(\alpha_k + Y_k^{(i)}\right), \end{aligned}$$

where $\psi(z) \equiv d(\ln\Gamma(z))/dz = \Gamma'(z)/\Gamma(z)$ is the so-called Digamma function. Hence,

$$\begin{aligned} &\frac{\partial^2}{\partial\alpha_k^2} \ln f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)} \,\middle|\, \underline{\alpha}\right) \\ &= N\left(\psi_1\left(\sum_{j=1}^{K+1}\alpha_j\right) - \psi_1\left(\sum_{j=1}^{K+1}\alpha_j + r\right) - \psi_1(\alpha_k)\right) + \sum_{i=1}^{N}\psi_1\left(\alpha_k + Y_k^{(i)}\right) \end{aligned}$$

and

$$\frac{\partial^2}{\partial\alpha_k\partial\alpha_l}\ln f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\,\middle|\,\underline{\alpha}\right)=N\left(\psi_1\left(\sum_{j=1}^{K+1}\alpha_j\right)-\psi_1\left(\sum_{j=1}^{K+1}\alpha_j+r\right)\right),$$

where $\psi_1(z)\equiv d\psi(z)/dz$ is the so-called Trigamma function. Unfortunately, the complicated nature of $E\left[\sum_{i=1}^{N}\psi_1\left(\alpha_k+Y_k^{(i)}\right)\right]$ makes the calculation of Jeffreys' hyperprior too difficult to pursue for Dirichlet-Multinomially distributed data.

### 5.1.3 Hyperprior $f(\underline{\alpha})\propto 1\Big/\left(\sum_{j=1}^{K+1}\alpha_j\right)^{(K+1)}$

Returning to the notation used in Chapter 4, let

$$\underline{\alpha}=\tau\underline{\lambda},$$

where $\tau>0$, $\underline{\lambda}>\underline{0}$ and $\sum_{j=1}^{K+1}\lambda_j=1$, or inversely, let

$$\begin{aligned}\lambda_j &= \frac{\alpha_j}{\sum_{k=1}^{K+1}\alpha_k}, \qquad j=1,\ldots,K,\\ \tau &= \sum_{j=1}^{K+1}\alpha_j\,.\end{aligned}$$

We will derive a hyperprior for $\underline{\alpha}$ by choosing priors for the transformed variables $\underline{\lambda}$ and $\tau$. Then

$$f(\underline{\alpha})=f(\underline{\lambda},\tau)|J|\,,$$

where

$$J=\begin{vmatrix}\frac{\partial\lambda_1}{\partial\alpha_1} & \frac{\partial\lambda_1}{\partial\alpha_2} & \cdots & \frac{\partial\lambda_1}{\partial\alpha_{K+1}}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial\lambda_K}{\partial\alpha_1} & \frac{\partial\lambda_K}{\partial\alpha_2} & \cdots & \frac{\partial\lambda_K}{\partial\alpha_{K+1}}\\ \frac{\partial\tau}{\partial\alpha_1} & \frac{\partial\tau}{\partial\alpha_2} & \cdots & \frac{\partial\tau}{\partial\alpha_{K+1}}\end{vmatrix}$$

is the Jacobian of the transformation. Specifically evaluating this determinant, by recursive use of the general result on p72 of [5], gives

$$\left(\frac{1}{\left(\sum_{j=1}^{K+1}\alpha_j\right)^2}\right)^{-(K+1)} J$$

$$= \begin{vmatrix} \sum_{j=1}^{K+1}\alpha_j - \alpha_1 & -\alpha_1 & \cdots & -\alpha_1 & -\alpha_1 \\ -\alpha_2 & \sum_{j=1}^{K+1}\alpha_j - \alpha_2 & \cdots & -\alpha_2 & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha_K & -\alpha_K & \cdots & \sum_{j=1}^{K+1}\alpha_j - \alpha_K & -\alpha_K \\ \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 \end{vmatrix}$$

$$= \begin{vmatrix} \sum_{j=1}^{K+1}\alpha_j & 0 & \cdots & 0 & 0 \\ -\alpha_2 & \sum_{j=1}^{K+1}\alpha_j - \alpha_2 & \cdots & -\alpha_2 & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha_K & -\alpha_K & \cdots & \sum_{j=1}^{K+1}\alpha_j - \alpha_K & -\alpha_K \\ \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 \end{vmatrix} +$$

$$\begin{vmatrix} -\alpha_1 & -\alpha_1 & \cdots & -\alpha_1 & -\alpha_1 \\ -\alpha_2 & \sum_{j=1}^{K+1}\alpha_j - \alpha_2 & \cdots & -\alpha_2 & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha_K & -\alpha_K & \cdots & \sum_{j=1}^{K+1}\alpha_j - \alpha_K & -\alpha_K \\ \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 \end{vmatrix}$$

$$= \left(\sum_{j=1}^{K+1}\alpha_j\right) \times$$

$$\begin{vmatrix} 1 & 0 & \cdots & 0 & 0 \\ -\alpha_2 & \sum_{j=1}^{K+1}\alpha_j - \alpha_2 & \cdots & -\alpha_2 & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha_K & -\alpha_K & \cdots & \sum_{j=1}^{K+1}\alpha_j - \alpha_K & -\alpha_K \\ \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 \end{vmatrix}$$

$$
\begin{aligned}
&= \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 \times \\
&\quad \begin{vmatrix}
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
-\alpha_K & -\alpha_K & \cdots & \sum_{j=1}^{K+1}\alpha_j - \alpha_K & -\alpha_K \\
\left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2
\end{vmatrix} \\
&\vdots \\
&= \left(\sum_{j=1}^{K+1}\alpha_j\right)^K \times \\
&\quad \begin{vmatrix}
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0 \\
\left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2 & \left(\sum_{j=1}^{K+1}\alpha_j\right)^2
\end{vmatrix} \\
&= \left(\sum_{j=1}^{K+1}\alpha_j\right)^{K+2} \begin{vmatrix}
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0 \\
1 & 1 & \cdots & 1 & 1
\end{vmatrix} \\
&= \left(\sum_{j=1}^{K+1}\alpha_j\right)^{K+2}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
J &= \left(\sum_{j=1}^{K+1}\alpha_j\right)^{-2(K+1)} \left(\sum_{j=1}^{K+1}\alpha_j\right)^{K+2} \\
&= \left(\sum_{j=1}^{K+1}\alpha_j\right)^{-K}. \qquad (5.2)
\end{aligned}
$$

An appropriate choice of noninformative prior for $\underline{\lambda}$, given $\tau$, would be

$$\begin{aligned} f(\underline{\lambda}|\tau) &= f(\underline{\lambda}) \\ &\sim \text{Dirichlet}(\underline{1}) . \end{aligned}$$

A common way of specifying an arbitrary and supposedly noninformative prior for the value of a positive parameter is to use the improper Jeffreys-Haldane density proportional to the inverse of that parameter. Hence, for $\tau$, a density of the form

$$f(\tau) \propto \frac{1}{\tau}$$

might be contemplated. Then the hyperprior for $\underline{\alpha}$ becomes

$$\begin{aligned} f(\underline{\alpha}) &= f(\underline{\lambda}, \tau)|J| \\ &= f(\underline{\lambda}|\tau) f(\tau)|J| \\ &\propto \frac{1}{\tau^{(K+1)}} , \end{aligned}$$

by (5.2). However, the following theorem shows that this also leads to an incoherent array of probabilities for $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$ and hence an undefined posterior for $\underline{\alpha}$.

**Theorem 5.1.3** *If*

$$f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right) \sim DMD(r, \underline{\alpha}), \qquad i = 1, \ldots, N+1 ,$$

*and*

$$f(\underline{\alpha}) \propto \frac{1}{\left(\sum_{j=1}^{K+1} \alpha_j\right)^{(K+1)}} ,$$

*then* $f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$ *is infinite and hence the posterior distribution for* $\underline{\alpha}$, $f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$, *is undefined, for every* $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$.

PROOF: For $j = 1, \ldots, K+1$, let $S_j = \sum_{i=1}^{N} Y_j^{(i)}$. Then

$$\begin{aligned} & f\left(\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) \\ = & \int_{\underline{\alpha}} \prod_{i=1}^{N} \left[ f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right) \right] f(\underline{\alpha}) d\underline{\alpha} \\ = & \int_{\underline{\alpha}} \prod_{i=1}^{N} \left[ \frac{r! \Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right) \prod_{j=1}^{K+1} \Gamma\left(\alpha_j + Y_j^{(i)}\right)}{\prod_{j=1}^{K+1} \left[Y_j^{(i)}!\right] \prod_{j=1}^{K+1} [\Gamma(\alpha_j)] \Gamma\left(\sum_{j=1}^{K+1} \alpha_j + r\right)} \right] \frac{1}{\left(\sum_{j=1}^{K+1} \alpha_j\right)^{(K+1)}} d\underline{\alpha} \\ \geq & \text{ const.} \int_{\underline{\alpha}} \frac{\prod_{j=1}^{K+1} \alpha_j^{S_j}}{\left(\sum_{j=1}^{K+1} \alpha_j + r - 1\right)^{(rN+K+1)}} d\underline{\alpha} \\ = & \infty , \end{aligned}$$

by the proof of Theorem 5.1.2. □

It appears that the search for suitable hyperpriors, $f(\underline{\alpha})$, is best concentrated on proper distributions, these being guaranteed to lead to a proper posterior for $\underline{\alpha}$. It may be of interest to study proper distributions that approximate improper noninformative priors. One such possibility will now be considered.

### 5.1.4 A Proper Approximation to $f(\underline{\alpha}) \propto 1/\left(\sum_{j=1}^{K+1} \alpha_j\right)^{(K+1)}$

Again let

$$\begin{aligned} f(\underline{\lambda}|\tau) &= f(\underline{\lambda}) \\ &\sim \text{Dirichlet}(\underline{1}) \end{aligned}$$

and consider approximating the improper Jeffreys-Haldane density $1/\tau$ by

$$f(\tau) = \frac{1}{\tau(\pi^2 + (\ln \tau)^2)}. \tag{5.3}$$

The density function given by (5.3) is a special case of the log-Cauchy density,

$$f(z) = \frac{\eta}{\pi z(\eta^2 + (\ln(z/\mu))^2)}, \qquad \eta = \ln(\nu/\mu) = \ln(\mu/\nu'),$$

where $\mu$ is the median, $\nu$ is the upper quartile and $\nu'$ is the lower quartile. Good [39, 40, 42] has advocated the use of various log-Cauchy distributions, in particular (5.3), for approximating $1/k$ when dealing with mixtures of symmetric Dirichlet distributions where $\underline{\alpha} = k\underline{1}$. Then

$$\begin{aligned} f(\underline{\alpha}) &= f(\underline{\lambda}, \tau)|J| \\ &= f(\underline{\lambda}|\tau) f(\tau)|J| \\ &\propto \frac{1}{\tau^{(K+1)}(\pi^2 + (\ln \tau)^2)}, \end{aligned} \tag{5.4}$$

by (5.2), and is a proper approximation to the noninformative hyperprior considered in the previous section.

The integrations involved in the calculation of the predictive probabilities using (5.1) with $f(\underline{\alpha})$ given by (5.4) do not lead to closed form solutions. Hence it was decided, as an example, to evaluate these probabilities numerically for the case $N + 1 = 4$, $r = 2$, $K + 1 = 3$. This was accomplished by generating a sample

of 100 000 $\underline{\alpha}$'s from the posterior distribution $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ and averaging $P\left(\underline{Y}^{(N+1)}=\underline{y}\middle|\underline{\alpha}\right)$ over these. This sample was produced using the 'rejection method' outlined in Smith and Gelfand [83]. The initial generation of points from (5.4) was easily done by generating $\lambda_1$, $\lambda_2$, $\lambda_3$ uniformly in the 2-dimensional simplex and using the inverse transformation method, as described in Pidd [75], to produce $\tau$ from (5.3). The calculations were performed in MATLAB VERSION 4.2C. Implementation of the rejection method requires knowledge of the value

$$M=\sup_{\underline{\alpha}>\underline{0}} f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\middle|\underline{\alpha}\right),$$

for each possible outcome $\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}$ in the set of conditioning histograms. For simplicity this set was taken to be $\mathcal{H}_P$, although the set of all possible conditioning histograms, $\mathcal{H}'$, would be a valid choice. To illustrate the manner in which these upper bounds were found, consider, for example, the histogram formed by

$$\left\{\underline{Y}^{(1)},\underline{Y}^{(2)},\underline{Y}^{(3)}\right\}=\left\{\begin{pmatrix}2\\0\\0\end{pmatrix},\begin{pmatrix}1\\0\\1\end{pmatrix},\begin{pmatrix}0\\1\\1\end{pmatrix}\right\}.$$

Then

$$\begin{aligned} g(\underline{\alpha}) &\equiv \prod_{i=1}^{N} f\left(\underline{Y}^{(i)}\middle|\underline{\alpha}\right) \\ &= \frac{4(\alpha_1+1)\alpha_1^2\alpha_2\alpha_3^2}{(\alpha_1+\alpha_2+\alpha_3)^3(\alpha_1+\alpha_2+\alpha_3+1)^3}. \end{aligned}$$

It can be easily verified that $\ln g(\underline{\alpha})$, and hence $g(\underline{\alpha})$, has no critical points for $\underline{\alpha}>\underline{0}$. To find an upper bound it therefore suffices to look at the behaviour of $g(\underline{\alpha})$ along the boundaries of the parameter space (including where some $\alpha_j$, $j=1,2,3$, tend to infinity). Note that

$$\begin{aligned} M &= \sup_{\underline{\alpha}>\underline{0}} g(\underline{\alpha}) \\ &= \sup_{\underline{\beta}>\underline{0}} \frac{4(\beta_1^2+1)\beta_1^4\beta_2^2\beta_3^4}{(\beta_1^2+\beta_2^2+\beta_3^2)^3(\beta_1^2+\beta_2^2+\beta_3^2+1)^3}, \end{aligned}$$

where $\alpha_j=\beta_j^2$, $j=1,2,3$. This transformation will simplify the development that follows. Letting

$$\begin{aligned}\beta_1 &= R\cos\theta\sin\phi \\ \beta_2 &= R\sin\theta\sin\phi \\ \beta_3 &= R\cos\phi\,,\end{aligned}$$

where $R > 0$, $0 \leq \theta, \phi \leq \pi/2$, we have

$$\begin{aligned}M &= \sup_{R,\theta,\phi} \frac{4(R^2\cos^2\theta\sin^2\phi+1)R^{10}\cos^4\theta\sin^2\theta\sin^6\phi\cos^4\phi}{R^6(R^2+1)^3} \\ &= \sup_{R,\theta,\phi} \frac{4R^4(R^2\cos^2\theta\sin^2\phi+1)}{(R^2+1)^3}\cos^4\theta\sin^2\theta\sin^6\phi\cos^4\phi \\ &\equiv \sup_{R,\theta,\phi} \mathfrak{g}(R,\theta,\phi)\,,\end{aligned}$$

say. Clearly, $g(\underline{\alpha})$ is zero along the $\alpha_1$, $\alpha_2$ and $\alpha_3\,(\phi = 0)$ axes, $\alpha_1\alpha_2\,(\phi = \pi/2)$, $\alpha_1\alpha_3\,(\theta = 0)$ and $\alpha_2\alpha_3\,(\theta = \pi/2)$ planes, and there is a removable discontinuity at the origin $(R = 0)$ where $g(\underline{\alpha}) \to 0$. Hence an upper bound for $g(\underline{\alpha})$ will be found by looking at the behaviour of $\mathfrak{g}(R,\theta,\phi)$ for large $R$. Now

$$\begin{aligned}G(\theta,\phi) &= \lim_{R\to\infty} \mathfrak{g}(R,\theta,\phi) \\ &= 4\cos^6\theta\sin^2\theta\sin^8\phi\cos^4\phi\,,\end{aligned}$$

so that

$$\begin{aligned}\frac{\partial G(\theta,\phi)}{\partial\theta} &= 8\sin^8\phi\cos^4\phi\cos^5\theta\sin\theta(1-4\sin^2\theta) \\ \frac{\partial G(\theta,\phi)}{\partial\phi} &= 16\cos^6\theta\sin^2\theta\sin^7\phi\cos^3\phi(2-3\sin^2\phi)\,.\end{aligned}$$

Setting the above two derivatives equal to zero and solving for $\theta, \phi \in (0, \pi/2)$ amounts to solving

$$\begin{aligned}\sin^2\theta &= \frac{1}{4} \qquad \left(\cos^2\theta = \frac{3}{4}\right) \\ \sin^2\phi &= \frac{2}{3} \qquad \left(\cos^2\phi = \frac{1}{3}\right)\,.\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{\theta,\phi}\lim_{R\to\infty}\mathfrak{g}(R,\theta,\phi) &= \sup_{\theta,\phi} G(\theta,\phi) \\ &= 4\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^2 \\ &= \frac{1}{108}\,.\end{aligned}$$

The uniform convergence of $g(R,\theta,\phi)$ to $G(\theta,\phi)$ can be easily verified, ensuring (see [79]) that

$$\lim_{R\to\infty}\sup_{\theta,\phi} \mathfrak{g}(R,\theta,\phi) = \sup_{\theta,\phi}\lim_{R\to\infty} \mathfrak{g}(R,\theta,\phi)\,.$$

Hence, for the histogram under consideration,

$$M = \frac{1}{108}\,.$$

It is interesting to note the direction in which the function $g(\underline{\alpha})$ approaches this upper bound. In terms of the original parameter $\underline{\alpha}$, we have

$$\begin{aligned}
\frac{\alpha_1}{\alpha_2} &= \frac{\beta_1^2}{\beta_2^2} \\
&= \frac{R^2\cos^2\theta\sin^2\phi}{R^2\sin^2\theta\sin^2\phi} \\
&= \frac{\cos^2\theta}{\sin^2\theta} \\
&= \frac{3/4}{1/4} \\
&= 3 \\
\Rightarrow \alpha_1 &= 3\alpha_2
\end{aligned}$$

and thus

$$\begin{aligned}
\frac{4\alpha_2}{\alpha_3} &= \frac{\alpha_1+\alpha_2}{\alpha_3} \\
&= \frac{R^2\cos^2\theta\sin^2\phi + R^2\sin^2\theta\sin^2\phi}{R^2\cos^2\phi} \\
&= \frac{\sin^2\phi}{\cos^2\phi} \\
&= \frac{2/3}{1/3} \\
&= 2 \\
\Rightarrow \alpha_3 &= 2\alpha_2\,.
\end{aligned}$$

Hence, $g(\underline{\alpha})$ approaches $1/108$ as $\omega\to\infty$ where $\underline{\alpha} \equiv (\alpha_1,\alpha_2,\alpha_3)^T = \omega(3,1,2)^T$. This direction corresponds to the relative powers of $\alpha_1$, $\alpha_2$ and $\alpha_3$ in the numerator of $g(\underline{\alpha})$, for large $\underline{\alpha}$. Clearly, for all of the histograms that correspond to a row (and possibly) column permutation of the current one, $1/108$ is the appropriate value

of $M$. This is because $g(\underline{\alpha}) \equiv \prod_{i=1}^{N} f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right)$ will only differ by an interchange of the roles of $\alpha_1$, $\alpha_2$ and $\alpha_3$ in the numerator, affecting the direction in which the function approaches its upper bound, but not the value of this bound.

It turns out that the value of $M$ for most histograms from the set $\mathcal{H}_P$ can be obtained by looking at

$$\lim_{\omega \to \infty} g\left(\omega(m,n,p)^T\right),$$

where $m$, $n$ and $p$ correspond to the relative powers of $\alpha_1$, $\alpha_2$ and $\alpha_3$ in the numerator of $g(\underline{\alpha})$, for large $\underline{\alpha}$. The exceptions are histograms of the same sort as that formed by

$$\left\{\underline{Y}^{(1)}, \underline{Y}^{(2)}, \underline{Y}^{(3)}\right\} = \left\{\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right\}$$

and the histogram formed by

$$\left\{\underline{Y}^{(1)}, \underline{Y}^{(2)}, \underline{Y}^{(3)}\right\} = \left\{\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}\right\}.$$

In the former case, $g(\underline{\alpha})$ has a critical point where $0 < \alpha_1, \alpha_2, \alpha_3 < 1$, leading to an upper bound of 0.0076 (4 d.p.). In the latter case, $g(\underline{\alpha})$ does not have a removable discontinuity at the origin and $M = 1/27$ is determined by looking at

$$\sup_{\theta,\phi} \lim_{R \to 0} \mathfrak{g}(R, \theta, \phi),$$

where the reparameterisation is as above.

The results of the numerical evaluation of the predictive probabilities using (5.1) with $f(\underline{\alpha})$ given by (5.4) are presented for the 29 relevant histograms in Table 5.1. In the first column of this table, the histograms in $\mathcal{H}_P$ are again identified by the $3 \times 3$ matrix whose columns represent the type outcomes of the 3 groups. Recall, also, the notation

$$p_{t,H} = P\left(\underline{Y}^{(N+1)} = \text{type } t \middle| (H(\underline{Y}_N) = H)\right), \qquad t = 1, \ldots, 6,$$

where types $1, \ldots, 6$ are

$$\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix},$$

respectively. The probabilities are quoted to 4 decimal places.

Given Theorem 4.3.3, we can use Theorem 2.4.1 to analyse the coherency of the strategy that has produced these predictive probabilities. This involves finding all of the $n$-cycles that are formed by histograms from $\mathcal{H}_P$ and determining whether or not they are satisfied for the probabilities in Table 5.1. This was accomplished by developing firstly an algorithm which takes the values $N+1$, $r$, $K+1$ and *size* as input, finds the set of histograms, $\mathcal{H}_P$, and finds all of the $n$-cycles formed by histograms from this set where $n = size$, and secondly an algorithm which outputs the minimum and maximum value of the *size*-cycle probability ratios for a given collection of probabilities. The algorithms were implemented in MAPLE V RELEASE 3 (see Appendix A, § A.5–A.6).

The results of the aforementioned algorithms may be summarised by quoting the overall minimum and maximum values of the $n$-cycle probability ratios, for all $n$. In fact, for the case at hand, $n$-cycles only exist for $3 \leq n \leq 12$. The overall minimum ratio of

$$0.956 \quad (3 \text{ d.p.})$$

corresponds to a 12-cycle formed by histograms $H2$, $H5$, $H23$, $H24$, $H29$, $H10$, $H9$, $H20$, $H18$, $H19$, $H13$ and $H11$, while the overall maximum ratio of

$$1.037 \quad (3 \text{ d.p.})$$

corresponds to a 9-cycle formed by histograms $H7$, $H8$, $H27$, $H15$, $H11$, $H13$, $H21$, $H20$ and $H9$. Given that the probabilities in Table 5.1 were obtained using approximate numerical methods, the above minimum and maximum $n$-cycle probability ratios are remarkably close to 1. They are sufficiently close to conclude that all of the $n$-cycles are satisfied. Practically speaking, this means that a strategy for estimating predictive probabilities that would employ (5.1) with

$$f(\underline{\alpha}) \propto \frac{1}{\tau^{(K+1)}(\pi^2 + (\ln \tau)^2)},$$

for all $H \in \mathcal{H}_P$, is indeed coherent when $N+1 = 4$, $r = 2$ and $K+1 = 3$. More importantly, it is coherent without the need for the concomitant assertion of zero probability of the first $N$ groups producing a positive histogram.

Two questions that naturally arise are whether or not this result extends (a) to other values of $N+1$, $r$ and $K+1$ for the same choice of $f(\underline{\alpha})$, and (b) to other

| Histogram, $H$ | $p_{1,H}$ | $p_{2,H}$ | $p_{3,H}$ | $p_{4,H}$ | $p_{5,H}$ | $p_{6,H}$ |
|---|---|---|---|---|---|---|
| $H1 = \begin{bmatrix} 220 \\ 001 \\ 001 \end{bmatrix}$ | 0.3522 | 0.1575 | 0.1580 | 0.1297 | 0.0717 | 0.1309 |
| $H2 = \begin{bmatrix} 211 \\ 010 \\ 001 \end{bmatrix}$ | 0.3510 | 0.1944 | 0.1945 | 0.0899 | 0.0801 | 0.0900 |
| $H3 = \begin{bmatrix} 210 \\ 011 \\ 001 \end{bmatrix}$ | 0.2389 | 0.2270 | 0.1515 | 0.1698 | 0.1197 | 0.0931 |
| $H4 = \begin{bmatrix} 210 \\ 010 \\ 002 \end{bmatrix}$ | 0.3003 | 0.1227 | 0.1581 | 0.1481 | 0.0864 | 0.1844 |
| $H5 = \begin{bmatrix} 210 \\ 002 \\ 010 \end{bmatrix}$ | 0.2999 | 0.1578 | 0.1225 | 0.1841 | 0.0866 | 0.1491 |
| $H6 = \begin{bmatrix} 210 \\ 001 \\ 011 \end{bmatrix}$ | 0.2390 | 0.1511 | 0.2273 | 0.0929 | 0.1195 | 0.1702 |
| $H7 = \begin{bmatrix} 200 \\ 021 \\ 001 \end{bmatrix}$ | 0.1843 | 0.1576 | 0.0864 | 0.3004 | 0.1224 | 0.1489 |
| $H8 = \begin{bmatrix} 200 \\ 020 \\ 002 \end{bmatrix}$ | 0.3078 | 0.0254 | 0.0255 | 0.3075 | 0.0255 | 0.3083 |
| $H9 = \begin{bmatrix} 200 \\ 011 \\ 011 \end{bmatrix}$ | 0.1459 | 0.1610 | 0.1606 | 0.1776 | 0.1777 | 0.1771 |
| $H10 = \begin{bmatrix} 200 \\ 010 \\ 012 \end{bmatrix}$ | 0.1838 | 0.0867 | 0.1578 | 0.1486 | 0.1228 | 0.3003 |

continued next page

| Histogram, $H$ | $p_{1,H}$ | $p_{2,H}$ | $p_{3,H}$ | $p_{4,H}$ | $p_{5,H}$ | $p_{6,H}$ |
|---|---|---|---|---|---|---|
| $H11 = \begin{bmatrix} 111 \\ 110 \\ 001 \end{bmatrix}$ | 0.2379 | 0.2453 | 0.1654 | 0.1482 | 0.1240 | 0.0791 |
| $H12 = \begin{bmatrix} 110 \\ 111 \\ 001 \end{bmatrix}$ | 0.1482 | 0.2459 | 0.1241 | 0.2377 | 0.1653 | 0.0788 |
| $H13 = \begin{bmatrix} 110 \\ 110 \\ 002 \end{bmatrix}$ | 0.1767 | 0.1776 | 0.1606 | 0.1775 | 0.1612 | 0.1465 |
| $H14 = \begin{bmatrix} 111 \\ 100 \\ 011 \end{bmatrix}$ | 0.2373 | 0.1651 | 0.2460 | 0.0786 | 0.1242 | 0.1487 |
| $H15 = \begin{bmatrix} 110 \\ 102 \\ 010 \end{bmatrix}$ | 0.1687 | 0.2275 | 0.1192 | 0.2401 | 0.1516 | 0.0929 |
| $H16 = \begin{bmatrix} 110 \\ 101 \\ 011 \end{bmatrix}$ | 0.1488 | 0.1850 | 0.1849 | 0.1479 | 0.1846 | 0.1489 |
| $H17 = \begin{bmatrix} 110 \\ 100 \\ 012 \end{bmatrix}$ | 0.1691 | 0.1197 | 0.2273 | 0.0930 | 0.1517 | 0.2392 |
| $H18 = \begin{bmatrix} 100 \\ 121 \\ 001 \end{bmatrix}$ | 0.0896 | 0.1940 | 0.0802 | 0.3512 | 0.1948 | 0.0902 |
| $H19 = \begin{bmatrix} 100 \\ 120 \\ 002 \end{bmatrix}$ | 0.1488 | 0.1225 | 0.0864 | 0.3003 | 0.1576 | 0.1843 |
| $H20 = \begin{bmatrix} 100 \\ 111 \\ 011 \end{bmatrix}$ | 0.0788 | 0.1653 | 0.1242 | 0.2377 | 0.2455 | 0.1485 |

continued next page

| Histogram, $H$ | $p_{1,H}$ | $p_{2,H}$ | $p_{3,H}$ | $p_{4,H}$ | $p_{5,H}$ | $p_{6,H}$ |
|---|---|---|---|---|---|---|
| $H21 = \begin{bmatrix} 100 \\ 110 \\ 012 \end{bmatrix}$ | 0.0925 | 0.1195 | 0.1515 | 0.1693 | 0.2276 | 0.2396 |
| $H22 = \begin{bmatrix} 100 \\ 100 \\ 022 \end{bmatrix}$ | 0.1303 | 0.0718 | 0.1579 | 0.1299 | 0.1577 | 0.3523 |
| $H23 = \begin{bmatrix} 110 \\ 002 \\ 110 \end{bmatrix}$ | 0.1771 | 0.1608 | 0.1771 | 0.1468 | 0.1609 | 0.1773 |
| $H24 = \begin{bmatrix} 110 \\ 001 \\ 111 \end{bmatrix}$ | 0.1483 | 0.1242 | 0.2461 | 0.0786 | 0.1649 | 0.2379 |
| $H25 = \begin{bmatrix} 100 \\ 022 \\ 100 \end{bmatrix}$ | 0.1305 | 0.1577 | 0.0719 | 0.3514 | 0.1579 | 0.1306 |
| $H26 = \begin{bmatrix} 100 \\ 021 \\ 101 \end{bmatrix}$ | 0.0924 | 0.1515 | 0.1193 | 0.2396 | 0.2277 | 0.1695 |
| $H27 = \begin{bmatrix} 100 \\ 020 \\ 102 \end{bmatrix}$ | 0.1487 | 0.0867 | 0.1227 | 0.1843 | 0.1577 | 0.3000 |
| $H28 = \begin{bmatrix} 100 \\ 011 \\ 111 \end{bmatrix}$ | 0.0787 | 0.1242 | 0.1654 | 0.1481 | 0.2458 | 0.2379 |
| $H29 = \begin{bmatrix} 100 \\ 010 \\ 112 \end{bmatrix}$ | 0.0900 | 0.0805 | 0.1942 | 0.0904 | 0.1943 | 0.3505 |

Table 5.1: Hierarchical Bayesian Predictive Probabilities for all $H \in H_P$, where $N+1 = 4$, $r = 2$, $K+1 = 3$ and $f(\underline{\alpha}) \propto 1/\left(\tau^{(K+1)}\left(\pi^2 + (\ln \tau)^2\right)\right)$

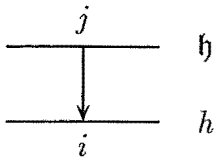
choices of hyperprior, $f(\underline{\alpha})$. The answer to both questions is yes, as the general theory of the next section will show.

## 5.2 General Priors

Many authors have studied the concept of coherency in relation to statistical inference, although their individual definitions of coherency often differ. Cornfield [20] and Freedman and Purves [31] introduced a formulation in which incoherence corresponds to an expected loss under some betting system which is uniformly positive when the expectation is taken over the possible values for the parameter or state of nature. This form of coherency has been studied by Heath and Sudderth [49], Lane and Sudderth [60], Regazzini [78] and Berti *et al.* [7]. It should be noted that such a definition is weaker than the one used in this thesis, since the loss is only guaranteed in the long run and not on every individual trial. Also, it makes little sense to act as though the underlying state of nature is a future observable. Despite differences in this detail, the general conclusion reached by these and other authors (Buehler [13], Lane [59], Lane and Sudderth [61], Gilio [36]) is the same: probabilities calculated with respect to Bayes' theorem for some proper prior distribution are coherent. The following theorem shows this to be true for the specific problem presently under consideration.

**Theorem 5.2.1** *Suppose you assert your predictive probabilities using (5.1) for all $H \in \mathcal{H}'$, the set of all possible conditioning histograms, to be strictly positive, where the hyperprior, $f(\underline{\alpha})$, leads to a defined posterior distribution for $\underline{\alpha}$. Then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them.*

PROOF: Consider a given $n$-cycle formed by histograms from $\mathcal{H}'$. Every histogram involved in the $n$-cycle appears as a subscript exactly once in the numerator and once in the denominator of the overall $n$-cycle probability ratio. Similarly, every type involved in the $n$-cycle appears an equal number of times as a subscript in this numerator and denominator. Hence, only the terms in (5.1) that depend simultaneously on the histogram *and* type will determine whether or not this $n$-cycle is satisfied. As we have already discovered, each occurrence




in the $n$-cycle contributes $p_{i,h}/p_{j,\mathfrak{h}}$ to the relevant ratio of probabilities, due to the fact that histogram $\mathfrak{h}$ contains one less type $j$ and one more type $i$ than does $h$. Let $\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}$ and $\underline{\mathcal{Y}}^{(1)},\ldots,\underline{\mathcal{Y}}^{(N)}$ be the outcomes of the first $N$ groups that form histograms $h$ and $\mathfrak{h}$, respectively. Then, looking only at the ratio of those terms in $p_{i,h}$ and $p_{j,\mathfrak{h}}$ that depend simultaneously on histogram and type gives

$$\frac{\int_{\underline{\alpha}} P\left(\underline{Y}^{(N+1)}=\text{type } i \,\middle|\, \underline{\alpha}\right)\prod_{i=1}^{N}\left[f\left(\underline{Y}^{(i)}\,\middle|\,\underline{\alpha}\right)\right]f(\underline{\alpha})d\underline{\alpha}}{\int_{\underline{\alpha}} P\left(\underline{Y}^{(N+1)}=\text{type } j \,\middle|\, \underline{\alpha}\right)\prod_{i=1}^{N}\left[f\left(\underline{\mathcal{Y}}^{(i)}\,\middle|\,\underline{\alpha}\right)\right]f(\underline{\alpha})d\underline{\alpha}}.$$

However, $\left\{\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)},\underline{Y}^{(N+1)}=\text{type } i\right\}=\left\{\underline{\mathcal{Y}}^{(1)},\ldots,\underline{\mathcal{Y}}^{(N)},\underline{Y}^{(N+1)}=\text{type } j\right\}$, so this ratio must equal 1. Hence the overall ratio of predictive probabilities equals 1 and the $n$-cycle is satisfied. The set $\mathcal{H}'$ is known to be linked and therefore Theorem 2.4.1 completes the proof. □

In light of Theorem 5.2.1, there is little point in producing and analysing *full* tables of probabilities such as Table 5.1, for different choices of $f(\underline{\alpha})$. However, it may be of interest to discuss possible hyperpriors and, recalling the results of Chapter 4, to see whether or not any of these distributions leads to the predictive probabilities using (5.1) being equal to

$$\frac{r!\Gamma(rN)\prod_{j=1}^{K+1}\Gamma\left(\sum_{i=1}^{N}Y_j^{(i)}+y_j\right)}{\prod_{j=1}^{K+1}[y_j!]\prod_{j=1}^{K+1}\left[\Gamma\left(\sum_{i=1}^{N}Y_j^{(i)}\right)\right]\Gamma(rN+r)}, \tag{5.5}$$

which is the empirical Bayes strategy (4.3) with $\hat{\alpha}_j=\sum_{i=1}^{N}Y_j^{(i)}$, $j=1,\ldots,K+1$, shown to be coherent in Theorem 4.3.4. For simplicity, this was checked by again considering the case $N+1=4$, $r=2$, $K+1=3$ and only for the histogram, $H$, formed by

$$\left\{\underline{Y}^{(1)},\underline{Y}^{(2)},\underline{Y}^{(3)}\right\}=\left\{\begin{pmatrix}2\\0\\0\end{pmatrix},\begin{pmatrix}1\\0\\1\end{pmatrix},\begin{pmatrix}0\\1\\1\end{pmatrix}\right\}.$$

For this histogram, $\left(\sum_{i=1}^{N}Y_1^{(i)},\sum_{i=1}^{N}Y_2^{(i)},\sum_{i=1}^{N}Y_3^{(i)}\right)^T=(3,1,2)^T$ and substituting type 1, ..., type 6 for $\underline{y}$ in (5.5) gives the probabilities

$$0.2857,\ 0.1429,\ 0.2857,\ 0.0476,\ 0.0952,\ 0.1429, \tag{5.6}$$

respectively, to 4 decimal places.

The following distributions were entertained as possible hyperpriors.

**No. 1:**

$$f(\underline{\alpha}) \propto \frac{1}{\tau^{(K+1)}(\pi^2 + (\ln \tau)^2)},$$

arising from

$$f(\underline{\lambda}|\tau) \sim \text{Dirichlet}(\underline{1}),$$
$$f(\tau) = \frac{1}{\tau(\pi^2 + (\ln \tau)^2)}.$$

**No. 2:** $f(\underline{\alpha})$ arising from

$$f(\underline{\lambda}|\tau) \sim \text{Dirichlet}(\underline{1}),$$

$$\begin{aligned} f(\rho) &= \begin{cases} 1, & \rho \in (0,1) \\ 0, & \text{otherwise} \end{cases} \\ &\sim \text{Uniform}(0,1), \end{aligned}$$

where $\rho \equiv 1/(1+\tau)$.

**No. 3:**

$$f(\underline{\alpha}) \propto \frac{\eta \exp^{-\eta\tau}}{\tau^K},$$

arising from

$$f(\underline{\lambda}|\tau) \sim \text{Dirichlet}(\underline{1}),$$

$$\begin{aligned} f(\tau) &= \eta \exp^{-\eta\tau} \\ &\sim \text{Exponential}(\eta). \end{aligned}$$

**No. 4:**

$$f(\underline{\alpha}) \propto \begin{cases} 1/(c\tau^K), & \tau \in (0,c) \\ 0, & \text{otherwise}, \end{cases}$$

arising from

$$f(\underline{\lambda}|\tau) \sim \text{Dirichlet}(\underline{1}),$$

$$\begin{aligned} f(\tau) &= \begin{cases} 1/c, & \tau \in (0,c) \\ 0, & \text{otherwise} \end{cases} \\ &\sim \text{Uniform}(0,c). \end{aligned}$$

**No. 5**: $f(\underline{\alpha})$ arising from

$$f(\alpha_j) \sim \text{Uniform}(0, c), \qquad j = 1, \ldots, K+1\,.$$

**No. 6**:

$$f(\underline{\alpha}) = \begin{cases} 1/10, & \underline{\alpha} = (4,1,1)^T \\ 1/10, & \underline{\alpha} = (1,4,1)^T \\ 1/10, & \underline{\alpha} = (1,1,4)^T \\ 1/10, & \underline{\alpha} = (3,2,1)^T \\ 1/10, & \underline{\alpha} = (3,1,2)^T \\ 1/10, & \underline{\alpha} = (2,3,1)^T \\ 1/10, & \underline{\alpha} = (2,1,3)^T \\ 1/10, & \underline{\alpha} = (1,3,2)^T \\ 1/10, & \underline{\alpha} = (1,2,3)^T \\ 1/10, & \underline{\alpha} = (2,2,2)^T \\ 0, & \text{otherwise}\,. \end{cases}$$

**No. 7**: $f(\underline{\alpha})$ arising from

$$\alpha_j \equiv k, \qquad j = 1, \ldots, K+1,$$

$$f(k) = \begin{cases} 1/2, & 0 < k < 1 \\ 1/(2k^2), & k \geq 1 \\ 0, & \text{otherwise}\,. \end{cases}$$

**No. 8**: $f(\underline{\alpha})$ arising from

$$\alpha_j \equiv k, \qquad j = 1, \ldots, K+1,$$

$$f(k) = \frac{1}{k(\pi^2 + (\ln k)^2)}\,.$$

**No. 9**: $f(\underline{\alpha})$ arising from

$$\alpha_j \equiv k, \qquad j = 1, \ldots, K+1,$$

$$f(k) = \frac{1}{\pi k(1 + (\ln k)^2)}\,.$$

**No. 10**: $f(\underline{\alpha})$ arising from

$$\alpha_j \equiv k, \qquad j = 1, \ldots, K+1,$$

$$f(k) = \frac{(K+1)\ln 7}{(7 + (K+1)k)(\ln(7 + (K+1)k))^2}\,.$$

**No. 11**: $f(\underline{\alpha})$ arising from

$$\alpha_j \equiv k, \qquad j = 1, \ldots, K+1,$$

$$f(k) \sim \text{Uniform}(0, c).$$

The first four of the above numbered distributions for $f(\underline{\alpha})$ are motivated by a desire to be noncommittal about the location of the mean parameters $\lambda_1, \ldots, \lambda_{K+1}$. The first, third and fourth each correspond to an example of the type of hyperprior structure suggested by Leonard [66]. The various accompanying densities for $\tau$ allow for a range in the available knowledge about this scale parameter, with the choice in No. 1 being the proper approximation already encountered to the noninformative density $1/\tau$. The choice of hyperprior No. 2 represents a generalisation of the approach involving reparameterisation taken in Lee and Sabavala [64]. Hyperprior No. 5 allows for more specific prior information regarding the parameter $\underline{\alpha}$. The discrete Uniform density No. 6 is included as a deliberate attempt to force the posterior for $\underline{\alpha}$ to give high weight to the vector $\left(\sum_{i=1}^{N} Y_1^{(i)}, \sum_{i=1}^{N} Y_2^{(i)}, \sum_{i=1}^{N} Y_3^{(i)}\right)^T$, perhaps leading to predictive probabilities similar to (5.6). Hyperpriors Nos. 7–11 all correspond to the assumption of a mixture of symmetric Dirichlet distributions for $\underline{\theta}^{(i)}$, $i = 1, \ldots, N+1$. This approach has been extensively used by Good [38, 39, 40, 41, 42] when the initial information concerning the $K+1$ categories is symmetrical. In fact, in [43] he formulates the Duns-Ockham hyper-razor as, "What can be done with fewer (hyper)parameters is done in vain with more". In all but the last of these five distributions the density $f(k)$ is one of the suggested approximations given in [39] to the improper Jeffreys-Haldane density $1/k$. (A simpler version of the proof of Theorem 5.1.3 shows that assuming $\alpha_j \equiv k$, $j = 1, \ldots, K+1$, and $f(k) \propto 1/k$, leads to an undefined posterior distribution for $\underline{\alpha}$.) The choice of $f(k)$ in hyperprior No. 11 allows for the incorporation of more specific prior information regarding the parameter $k$.

With the exception of density No. 6, the integrations involved in the calculation of the predictive probabilities using (5.1), with any of the above numbered hyperpriors, do not lead to closed form solutions. Hence these probabilities were evaluated numerically by generating a sample of 200 000 $\underline{\alpha}$'s from the posterior distribution $f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$ and averaging $P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{\alpha}\right)$ over these. Again, this sample was produced using the rejection method outlined in Smith and Gelfand [83].

The calculations were performed in MATLAB VERSION 4.2C and the probabilities are quoted in Table 5.2, along with the probabilities in (5.6), to 4 decimal places.

| Hyperprior, $f(\underline{\alpha})$, | $p_{1,H}$ | $p_{2,H}$ | $p_{3,H}$ | $p_{4,H}$ | $p_{5,H}$ | $p_{6,H}$ |
|---|---|---|---|---|---|---|
| No. 1 | 0.2389 | 0.1517 | 0.2274 | 0.0929 | 0.1198 | 0.1693 |
| No. 2 | 0.2586 | 0.1226 | 0.1838 | 0.1221 | 0.1037 | 0.2092 |
| No. 3, $\eta = 1/6$ | 0.2435 | 0.1400 | 0.2107 | 0.1037 | 0.1163 | 0.1857 |
| No. 3, $\eta = 1/60$ | 0.2266 | 0.1686 | 0.2528 | 0.0759 | 0.1299 | 0.1462 |
| No. 4, $c = 10$ | 0.2428 | 0.1411 | 0.2116 | 0.1031 | 0.1174 | 0.1841 |
| No. 4, $c = 100$ | 0.2257 | 0.1702 | 0.2554 | 0.0741 | 0.1307 | 0.1438 |
| No. 5, $c = 10$ | 0.1877 | 0.1782 | 0.2414 | 0.0928 | 0.1532 | 0.1468 |
| No. 5, $c = 100$ | 0.1806 | 0.1897 | 0.2586 | 0.0807 | 0.1581 | 0.1322 |
| No. 6 | 0.2183 | 0.1653 | 0.2093 | 0.1060 | 0.1384 | 0.1626 |
| No. 7 | 0.1610 | 0.1723 | 0.1723 | 0.1610 | 0.1723 | 0.1610 |
| No. 8 | 0.1379 | 0.1954 | 0.1954 | 0.1379 | 0.1954 | 0.1379 |
| No. 9 | 0.1548 | 0.1785 | 0.1785 | 0.1548 | 0.1785 | 0.1548 |
| No. 10 | 0.1272 | 0.2061 | 0.2061 | 0.1272 | 0.2061 | 0.1272 |
| No. 11, $c = 10$ | 0.1317 | 0.2016 | 0.2016 | 0.1317 | 0.2016 | 0.1317 |
| No. 11, $c = 100$ | 0.1148 | 0.2186 | 0.2186 | 0.1148 | 0.2186 | 0.1148 |
| $\alpha_j \equiv \sum_{i=1}^{N} Y_j^{(i)}$ | 0.2857 | 0.1429 | 0.2857 | 0.0476 | 0.0952 | 0.1429 |

Table 5.2: Hierarchical Bayesian Predictive Probabilities Conditioning on Histogram $H = \begin{bmatrix} 210 \\ 001 \\ 011 \end{bmatrix}$ ($N+1=4$, $r=2$, $K+1=3$) for Different Choices of Hyperprior, $f(\underline{\alpha})$

Clearly, none of the other rows of probabilities in Table 5.2 matches the last one. In fact, it would be impossible for any hyperprior given by a mixture of symmetric Dirichlet distributions to produce the predictive probabilities in (5.6). This is because the posterior $f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$ will still have all of its mass concentrated on points where $\alpha_1 = \cdots = \alpha_{K+1}$, leading to probabilities of the form

$$p_{1,H},\ p_{2,H},\ p_{3,H} = p_{2,H},\ p_{4,H} = p_{1,H},\ p_{5,H} = p_{2,H},\ p_{6,H} = p_{1,H}$$

for the case $r = 2$, $K + 1 = 3$ presently under consideration. Not surprisingly, therefore, calculation of the squared distance of each set of probabilities in the

rows of Table 5.2 from those in the last row produces a larger value for all of the hyperpriors corresponding to a mixture of symmetric Dirichlet distributions than for any of the other hyperpriors considered. The minimum squared distance of 0.0072 (4 d.p.) occured for both hyperprior No. 3 with $\eta = 1/60$ and hyperprior No. 4 with $c = 100$.

It is perhaps interesting to note that the discrete Uniform hyperprior No. 6, for which exact calculation of the posterior $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ is possible, does give maximum posterior probability to $\underline{\alpha} = \left(\sum_{i=1}^{N} Y_1^{(i)}, \sum_{i=1}^{N} Y_2^{(i)}, \sum_{i=1}^{N} Y_3^{(i)}\right)^T$, for all $H(\underline{Y}_N) \in \mathcal{H}_P$. Obviously if it were to give posterior probability 1 to this particular $\underline{\alpha}$ vector, the predictive probabilities using (5.1) would be exactly those in (5.6). This motivates the following section.

### 5.2.1 The Gibbs Sampler

Rather than trying to guess a distribution for $f(\underline{\alpha})$ that might lead to

$$f(\underline{\alpha}\,|\,H(\underline{Y}_N) \in \mathcal{H}_P) = \begin{cases} 1, & \alpha_j = \sum_{i=1}^{N} Y_j^{(i)},\ j = 1,\ldots,K+1, \\ 0, & \text{otherwise}, \end{cases} \tag{5.7}$$

it may be possible to use the Gibbs sampler, somewhat unconventionally, to find such a density. An algorithm for extracting marginal distributions from the full conditional distributions was formally introduced as the Gibbs sampler in Geman and Geman [35], although its roots date at least to Metropolis *et al.* [71], with further development by Hastings [48]. More recently, Gelfand and Smith [34] generated new interest in the technique by revealing its potential in a wide variety of statistical problems. An excellent explanation of the algorithm may be found in Casella and George [14].

Given $f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\middle|\underline{\alpha}\right)$ and $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$, it would be more common to use the Gibbs sampler to find the marginal distribution $f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$. However, it is also possible to ascertain $f(\underline{\alpha})$ using this approach. Taking

$$\begin{aligned} f\left(\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\middle|\underline{\alpha}\right) &= \prod_{i=1}^{N} f\left(\underline{Y}^{(i)}\middle|\underline{\alpha}\right) \\ &\sim \prod_{i=1}^{N} \mathrm{DMD}(r,\underline{\alpha}), \end{aligned} \tag{5.8}$$

$f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ as in (5.7) and contemplating implementation of the Gibbs sampler, one problem becomes immediately apparent. The generation of a histogram from (5.8) is not guaranteed to produce a positive histogram and therefore generation of an $\underline{\alpha}$ from $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ at the next step may not be defined. Repeatedly sampling from (5.8) until a positive histogram is obtained would obviously affect the convergence properties of the Gibbs sequence. Therefore a full specification of the posterior $f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right)$ is required.

A possible extension of (5.7) to deal with the case where $H(\underline{Y}_N) \notin \mathcal{H}_P$ would be to give probabilities summing to 1 to the $\underline{\alpha}$ vectors with integer-valued components 'nearest' to $\sum_{i=1}^{N} Y_j^{(i)}$ for $j = 1,\ldots,K+1$. For example, consider again the case $N+1=4$, $r=2$, $K+1=3$ and let $S_j = \sum_{i=1}^{N} Y_j^{(i)}$, $j=1,\ldots,3$. Then a possible conditional distribution is given by

$$
\begin{aligned}
& f\left(\underline{\alpha}\middle|\underline{Y}^{(1)},\ldots,\underline{Y}^{(N)}\right) \\
= & \begin{cases}
\left.\begin{array}{ll} 1, & \alpha_j = S_j,\ j=1,\ldots,3, \\ 0, & \text{otherwise} \end{array}\right\} \text{if } H(\underline{Y}_N) \in \mathcal{H}_P, \\
\left.\begin{array}{ll} 1, & \underline{\alpha} = (4,1,1)^T \\ 0, & \text{otherwise} \end{array}\right\} \text{if } (S_1,S_2,S_3)^T \in \{(6,0,0)^T,(5,1,0)^T\}, \\
\left.\begin{array}{ll} 1/2, & \underline{\alpha} = (4,1,1)^T \\ 1/2, & \underline{\alpha} = (3,2,1)^T \\ 0, & \text{otherwise} \end{array}\right\} \text{if } (S_1,S_2,S_3)^T = (4,2,0)^T, \\
\left.\begin{array}{ll} 1/2, & \underline{\alpha} = (3,2,1)^T \\ 1/2, & \underline{\alpha} = (2,3,1)^T \\ 0, & \text{otherwise} \end{array}\right\} \text{if } (S_1,S_2,S_3)^T = (3,3,0)^T, \\
\quad \textit{etc.}
\end{cases}
\end{aligned}
\tag{5.9}
$$

Attempts at running the Gibbs sampler with (5.8) and (5.9) failed to converge, suggesting that there does not exist a hyperprior, $f(\underline{\alpha})$, for which the posterior is (5.9).

## 5.3 An Alternative Model

Suppose a slightly different assertion was made regarding the nature of the $N+1$ groups of interest, namely that

$$\underline{\theta}^{(1)} = \cdots = \underline{\theta}^{(N+1)} \equiv \underline{\theta}\,,$$

say. Then

$$\begin{aligned} f\left(\underline{Y}^{(i)} \middle| \underline{\theta}\right) &= \frac{r!}{\prod_{j=1}^{K+1} Y_j^{(i)}!} \prod_{j=1}^{K+1} [\theta_j]^{Y_j^{(i)}} \\ &\sim \text{Multinomial}(r, \underline{\theta})\,, \qquad i = 1, \ldots, N+1\,, \end{aligned}$$

and

$$\begin{aligned} f\left(\underline{\theta} \,|\, \underline{\alpha}\right) &= \frac{\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right)}{\prod_{j=1}^{K+1} \Gamma(\alpha_j)} \prod_{j=1}^{K+1} [\theta_j]^{\alpha_j - 1} \\ &\sim \text{Dirichlet}(\underline{\alpha})\,, \end{aligned}$$

so that

$$\begin{aligned} f\left(\underline{Y}^{(i)} \middle| \underline{\alpha}\right) &= \frac{r!\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right) \prod_{j=1}^{K+1} \Gamma\left(\alpha_j + Y_j^{(i)}\right)}{\prod_{j=1}^{K+1} \left[Y_j^{(i)}!\right] \prod_{j=1}^{K+1} [\Gamma(\alpha_j)] \Gamma\left(\sum_{j=1}^{K+1} \alpha_j + r\right)} \\ &\sim \text{DMD}(r, \underline{\alpha}), \qquad i = 1, \ldots, N+1\,, \end{aligned}$$

where $\underline{\alpha} > \underline{0}$. The $\underline{Y}^{(i)}$ vectors, $i = 1, \ldots, N+1$, are presumed independent conditional upon $\underline{\theta}$, a structure the subjectivist would recognise as motivating an exchangeable distribution over the $\underline{Y}^{(i)}$. Thus the exact equation structure that applied in previous coherency analyses would remain valid for this new situation.

Many authors have looked at ways of estimating a multinomial probability vector, $\theta \equiv (\theta_1, \ldots, \theta_J)^T$, in the presence of observed data $\underline{n} \equiv (n_1, \ldots, n_J)^T$ where $\sum_{j=1}^{J} n_j = n$. Frequently considered are estimators of the form

$$\hat{\theta}_j = \beta \frac{n_j}{n} + (1 - \beta)\lambda_j, \qquad j = 1, \ldots, J\,, \tag{5.10}$$

where $\underline{\lambda}$ is a fixed point in the $K$-dimensional simplex and $0 \leq \beta \leq 1$. Another way of writing (5.10) is

$$\hat{\theta}_j = \frac{n_j + \tau\lambda_j}{n + \tau}, \qquad j = 1, \ldots, J\,, \tag{5.11}$$

where $\beta = n/(n+\tau)$. Expressed in this form, (5.11) is easily recognisable as the posterior mean of $\theta_j$ resulting from a Dirichlet prior with parameter $\tau\underline{\lambda}$. Often of interest is the 'equiprobable' case $\lambda_j = 1/J\ \forall j$, corresponding to the assumption of a symmetric Dirichlet prior for $\underline{\theta}$. Let $\tau\underline{\lambda} = k\underline{1}$. The choices for $k$ that have been proposed in the literature may be divided into two types — *a priori* values and empirically determined values. The *a priori* choices of $k$ include: $k = 0$ (corresponding to MLE of $\underline{\theta}$); $k = 1$ (Lidstone [69] generalising Laplace's law of succession); $k = 1/2$ (Jeffreys [52] using his invariance theory); and $k = 1/J$ (Perks [74]). Methods which use the data to estimate $k$ may be based on the repeat rate (see Chapter 4, §4.5.4) or on MLE of $k$. Trybula [85] showed that the minimax estimator of $\underline{\theta}$ under quadratic loss corresponds to $k = \sqrt{n}/J$, while Good [38, 40], Fienberg and Holland [28], Stone [84] and Leonard [67] all propose values for $k$ that depend on the data through

$$X^2 = \frac{\sum_{j=1}^{J}(n_j - n/J)^2}{n/J},$$

the $\chi^2$ statistic for testing $\theta_j = 1/J$, $j = 1, \ldots, J$.

In a hierarchical Bayesian approach where $\tau$ is given a hyperprior with density $f(\tau)$, the posterior probabilities are

$$E(\theta_j|\underline{n}) = \frac{n_j + \tau_0\lambda_j}{n + \tau_0},$$

where

$$\tau_0 = \frac{\int_0^\infty \frac{\tau}{n+\tau} f(\underline{n}|\tau\underline{\lambda}) f(\tau) d\tau}{\int_0^\infty \frac{1}{n+\tau} f(\underline{n}|\tau\underline{\lambda}) f(\tau) d\tau}$$

and

$$\begin{aligned} f(\underline{n}|\tau\underline{\lambda}) &= \frac{n!\Gamma(\tau)\prod_{j=1}^{J}\Gamma(\tau\lambda_j + n_j)}{\prod_{j=1}^{J}[n_j!]\prod_{j=1}^{J}[\Gamma(\tau\lambda_j)]\Gamma(\tau+n)} \\ &\sim \mathrm{DMD}(n, \tau\underline{\lambda}) \end{aligned}$$

(see Bishop *et al.* [8], p406). The expressions $\tau\lambda_j$ and $\tau_0\lambda_j$ are generally referred to as 'flattening constants' because they smooth the raw data proportions.

Returning to the problem at hand, the data, $\underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}$, from the first $N$ groups may be manipulated to fit this alternative model in one of two ways.

Case (a) Let $J = {}^{r+K}C_K$ and imagine redefining the categories to correspond to the different types of possible outcome for $\underline{Y}^{(i)}$. That is, let

$$\theta_t = P\left(\underline{Y}^{(i)} = \text{type } j \,\middle|\, \underline{\theta}\right), \qquad j = 1, \ldots, {}^{r+K}C_K,\ i = 1, \ldots, N+1,$$

and $\underline{n} = (x_1, \ldots, x_{r+\kappa C_K})^T$, where $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$. Then if the prior for $\underline{\theta}$ is taken to be Dirichlet($\underline{c}$), the posterior distribution $f\left(\underline{\theta} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$ is Dirichlet($\underline{n} + \underline{c}$) and we have

$$\begin{aligned} & P\left(\underline{Y}^{(N+1)} = \text{type } t \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) \\ = & \int_{\underline{\theta}} P\left(\underline{Y}^{(N+1)} = \text{type } t \middle| \underline{\theta}\right) f\left(\underline{\theta} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) d\underline{\theta} \\ = & \int_{\underline{\theta}} \theta_t f\left(\underline{\theta} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) d\underline{\theta} \\ = & E\left(\theta_t \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) \\ = & \frac{x_t + c_t}{N + \sum_{s=1}^{r+K C_K} c_s}. \end{aligned}$$

This is exactly the form suggested for the predictive probabilities in (3.6) of Chapter 3, and proven to define a coherent estimation strategy in Theorem 3.2.3.

Case (b) Let $J = K + 1$ and imagine pooling the samples from the $N$ groups. That is, let

$$\theta_j = P\left(\begin{array}{c} \text{an item from the } i^{\text{th}} \text{ group} \\ \text{is classified in category } j \end{array}\right), \qquad j = 1, \ldots, K+1,\ i = 1, \ldots, N+1,$$

and $\underline{n} = \left(\sum_{i=1}^N Y_1^{(i)}, \ldots, \sum_{i=1}^N Y_{K+1}^{(i)}\right)^T$. Then if the prior for $\underline{\theta}$ is taken to be Dirichlet($\underline{c}$), the posterior distribution $f\left(\underline{\theta} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right)$ is Dirichlet($\underline{n} + \underline{c}$). Letting $\underline{\alpha} = \underline{n} + \underline{c}$ and $\tau = \sum_{j=1}^{K+1} \alpha_j$, the posterior mean, $E\left[\theta_j \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right]$ is given by

$$\frac{\alpha_j}{\tau} = \frac{\sum_{i=1}^N Y_j^{(i)} + c_j}{rN + \sum_{k=1}^{K+1} c_k}, \qquad j = 1, \ldots, K+1, . \tag{5.12}$$

Also, we have

$$\begin{aligned} & P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) \\ = & \int_{\underline{\theta}} P\left(\underline{Y}^{(N+1)} = \underline{y} \middle| \underline{\theta}\right) f\left(\underline{\theta} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) d\underline{\theta} \\ = & \frac{r!\Gamma\left(\sum_{j=1}^{K+1} \alpha_j\right) \prod_{j=1}^{K+1} \Gamma\left(\alpha_j + y_j\right)}{\prod_{j=1}^{K+1} [y_j!] \prod_{j=1}^{K+1} [\Gamma(\alpha_j)] \Gamma\left(\sum_{j=1}^{K+1} \alpha_j + r\right)} \\ \sim & \text{DMD}(r, \underline{\alpha}), \end{aligned} \tag{5.13}$$

where $\alpha_j = \sum_{i=1}^N Y_j^{(i)} + c_j$, $j = 1, \ldots, K+1$. The reader's attention is drawn to the similarity between (5.12) and expression (4.82) of Chapter 4, and to the fact that

(5.13) is shown in Theorem 4.5.3 to be a coherent strategy for estimation of the predictive probabilities.

It therefore seems possible that all of the strategies in Chapters 3 and 4 that were found to be coherent for estimating your predictive probabilities may be so only because they correspond to a hierarchical Bayesian method under the assertion of a slightly different and more restrictive model, for which the coherency induced equations still happen to apply.

This completes a study of the coherency of various methods and strategies for specifying predictive probabilities. The main findings are now summarised in Chapter 6.

# Chapter 6

# Summary

Throughout this thesis attention has focussed on the estimation of the predictive probabilities

$$P\left(\underline{Y}^{(N+1)} = \underline{y} \,\middle|\, \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right), \tag{6.1}$$

where $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, is a vector of category counts from the $i^{\text{th}}$ of $N+1$ groups that are regarded exchangeably. Formulation of the problem in the framework of subjective statistics shows that any such procedure is governed by a system of homogeneous equations induced by the requirement of coherency. The variables in these equations represent your previsions for only those outcomes of the $N+1$ groups including among them $N$ groups that form a conditioning histogram for which you are prepared to assert probabilities as in (6.1). A feature that conditioning histograms are linked has been found to play a central role in characterising types of solutions to this system of equations. If the system is linked and all of your asserted predictive probabilities are strictly positive, then either there is a solution in which all of the variables may be expressed as nonzero multiples of any one of them, or only the trivial solution exists. In the former case, coherency allows you some freedom in the way in which you may assert your prevision for outcomes of the $N+1$ groups. However if only the trivial solution exists, the conclusion is that you must give zero probability to exactly those situations in which your specified predictive probabilities would be used! Therefore it is important to be able to determine conditions under which this is true for a given system of equations.

To formalise a solution to this issue, the concept of $n$-cycles has been introduced and developed in Chapter 2, culminating in Theorem 2.4.1. The $n$-cycle structures

that exist among the conditioning histograms are shown to be crucial in determining the existence of a nontrivial solution to a system of equations. It is worth commenting that although $n$-cycles surely exist where $n$ is quite a large number, it has been sufficient to consider only 3-cycles to prove results in all cases studied (see, for example, the proof of Theorem 4.3.6). Whether this is true more generally is a possible direction for future research and may suggest useful refinements to the definition of an $n$-cycle.

Conceptually, the $n$-cycle approach could be used to find all of the predictive probabilities that allow a nonzero solution to the system of coherency induced equations for a given problem, without having to solve the full system. However, in practice, even this seems to be computationally unfeasible or to produce expressions that are too complex to be of interest. The real strength of the $n$-cycle theory lies in its immediate applicability to the more likely scenario of analysing the coherency implications of a *specified* collection of predictive probabilities or strategy for estimating them. This has been illustrated in Chapters 3, 4 and 5 with the following main findings.

- Suppose you assert your predictive probabilities as

$$P\left(\underline{Y}^{(N+1)} = \text{type } t \,\middle|\, (H(\underline{Y}_N) = H)\right) = \frac{x_t + c_t}{N + \sum_{s=1}^{r+K C_K} c_s}, \quad t = 1, \ldots, {}^{r+K}C_K\,, \tag{6.2}$$

  for all $H \in \mathcal{H}_{SP}$, the set of strictly positive histograms, where $c_t \geq 0$, $c_t \in \mathbb{R}$, $t = 1, \ldots, {}^{r+K}C_K$, and $x_t$, $t = 1, \ldots, {}^{r+K}C_K$, denotes the number of $\underline{Y}^{(i)}$, $i = 1, \ldots, N$, of type $t$ in $H$. Then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them. On the contrary, if $c_t = 0\ \forall t$, amounting to a straight-out frequency mimicking approach, the *systematic* strategy to use (6.2) (given any strictly positive histogram) for *any* size of $N$ is only coherent along with the assertion

$$\sum_{H \in \mathcal{H}_{SP}} P(H(\underline{Y}_N) = H) = 0$$

  for every value of $N$. In other words, you must give zero probability to ever using your strategy. If $c_t = c > 0\ \forall t$, the systematic strategy to use (6.2) for any size of $N$ does not require this restrictive concomitant assertion. It

is summised that further research could prove that it is also not required if, more generally, $c_t > 0 \ \forall t$.

- Suppose you assert your predictive probabilities in (6.1) using a DMD$(r, \underline{\hat{\alpha}})$ with

$$\hat{\alpha}_j = \sum_{i=1}^{N} Y_j^{(i)} + c_j, \qquad j = 1, \ldots, K+1,$$

for all $H \in \mathcal{H}_P$, the set of positive histograms, where $c_j \geq 0, c_j \in \mathbb{R}$, $j = 1, \ldots, K+1$. Then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them. However, when substituted into the Dirichlet-Multinomial distribution, many other empirical Bayes estimates, $\underline{\hat{\alpha}}$, (due to Mosimann, Brier and others) give strategies that must be asserted with zero probability of being used in order to be coherent. Note that this last statement applies to a problem with particular values of $N+1, r$ and $K+1$, without requiring that the strategy be employed for every value of $N$. Further work might extend the range of methods considered for the empirical Bayes estimation of the parameter $\underline{\alpha}$ and investigate the assumption of an alternative distribution to the Dirichlet.

- Suppose you assert your predictive probabilities in (6.1) using

$$\int_{\underline{\alpha}} \text{DMD}(r, \underline{\alpha}) f\left(\underline{\alpha} \middle| \underline{Y}^{(1)}, \ldots, \underline{Y}^{(N)}\right) d\underline{\alpha}$$

for all $H \in \mathcal{H}'$, the set of all possible conditioning histograms, where the hyperprior, $f(\underline{\alpha})$, leads to a defined posterior distribution for $\underline{\alpha}$. Then the system of equations generated has a solution in which all of the variables may be expressed as nonzero multiples of any one of them. Again, future research could involve studying the assumption of an alternative distribution to the Dirichlet.

It has been pointed out in Chapter 5, §5.3 that the coherent strategies for estimating your predictive probabilities that are outlined in the first two points above may only be coherent because they correspond to a hierarchical Bayesian method under a different and more restrictive assertion that the $N+1$ groups are identical in nature. Does there exist any non-Bayesian, coherent strategy for the estimation of

the predictive probabilities in (6.1) that may be asserted with nonzero probability of being used and that truly applies to the general problem we have described, rather than merely to its restricted form? If so, the investigations conducted in this thesis have not revealed it and particularly those statisticians who do not adhere to the Bayesian school of thought must continue to address this question.

# Acknowledgements

I would like to extend my gratitude to my supervisor, Dr Frank Lad, for his guidance in the development of this research and for his continual support and encouragement. Comments on drafts of this manuscript have been invaluable.

I am also indebted to my associate supervisor, Professor John Deely, whose knowledge and experience in the field have been a great resource.

Of the departmental staff, Dr Neil Watson deserves special mention for his willingness to help and his ability to provide rigorous arguments. Fellow graduate students Dr Andrew Hill, Philip Schlüter, Chris Stephens and Julian Visch are thanked for useful discussions on matters of mathematics, statistics, motivation and typesetting, respectively.

Geoff M$^{c}$Ilraith is acknowledged for his friendship and for the development of a prototype computer program in C.

Most importantly, this thesis may never have made it to completion without the unwavering support of my family and their belief in me. Love to my sisters Sandra, Susan and Gaynor, my father, Harry, and especially to my mother, Valerie.

The financial support of a Ministry of Research, Science and Technology Postgraduate Study Award, a University Grants Committee Postgraduate Scholarship and a William and Ina Cartwright Scholarship are all gratefully acknowledged.

# References


[1] Aitchison, J. and Shen, S.M. (1980), "Logistic-Normal Distributions: Some Properties and Uses", *Biometrika* **67**, 261–272

[2] Aitchison, J. (1985), "A General Class of Distributions on the Simplex", *Journal of the Royal Statistical Society*, Series B **47**, 136–146

[3] Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London: Chapman and Hall

[4] Ahlfors, L.V. (1953), *Complex Analysis*, New York: McGraw-Hill

[5] Anton, H. (1984), *Elementary Linear Algebra*, 4th edition, New York: Wiley

[6] Attwell, D.N. and Smith, J.Q. (1991), "A Bayesian Forecasting Model for Sequential Bidding", *Journal of Forecasting* **10**, 565–577

[7] Berti, P., Regazzini, E. and Rigo, P. (1991), "Coherent Statistical Inference and Bayes Theorem", *The Annals of Statistics* **19**, 366–381

[8] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975), *Discrete Multivariate Analysis*, Cambridge, Massachusetts: The Massachusetts Institute of Technology Press

[9] Block, D. and Watson, G. (1967), "A Bayesian Study of the Multinomial Distribution", *The Annals of Mathematical Statistics* **38**, 1423–1435

[10] Borel, E. (1964), "Apropos of a Treatise on Probability", English translation (by H.E. Smokler) in *Studies in Subjective Probability*, (Editors: H.E. Kyburg and H.E. Smokler), New York: Wiley

[11] Brier, S.S. (1979), *Categorical Data Models for Complex Sampling Schemes*, Ph.D. thesis, University of Minnesota, Ann Arbor, Michigan: University Microfilms International

[12] Brier, S.S. (1980), "Analysis of Contingency Tables Under Cluster Sampling", *Biometrika* **67**, 591–596

[13] Buehler, R.J. (1976), "Coherent Preferences", *The Annals of Statistics* **4**, 1051–1064

[14] Casella, G. and George, E.I. (1992), "Explaining the Gibbs Sampler", *The American Statistician* **46**, 167–174

[15] Chandon, J-L. J. (1976), *A Comparative Study of Media Exposure Models*, Ph.D. thesis, Northwestern University

[16] Chatfield, C. and Goodhardt, G.J. (1970), "The Beta-Binomial Model for Consumer Purchasing Behaviour", *Applied Statistics* **19**, 240–250

[17] Chuang, C. and Cox, C. (1985), "Pseudo Maximum Likelihood Estimation for the Dirichlet-Multinomial Distribution", *Communications in Statistics - Theory and Methods* **14**, 2293–2311

[18] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling from Contingency Tables", *Journal of the American Statistical Association* **71**, 665–670

[19] Connor, R.J. and Mosimann, J.E. (1969), "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution", *Journal of the American Statistical Association* **64**, 194–206

[20] Cornfield, J. (1969), "The Bayesian Outlook and its Applications", *Biometrics* **25**, 617–642

[21] Danaher, P.J. (1988), "Parameter Estimation for the Dirichlet-Multinomial Distribution Using Supplementary Beta-Binomial Data", *Communications in Statistics - Theory and Methods* **17**, 1777–1788

[22] de Finetti, B. (1964), "Foresight: Its Logical Laws, its Subjective Sources", English translation (by H.E. Kyburg) in *Studies in Subjective Probability*, (Editors: H.E. Kyburg and H.E. Smokler), New York: Wiley

[23] de Finetti, B. (1972), *Probability, Induction and Statistics*, London: Wiley

[24] de Finetti, B. (1974, 1975) *Theory of Probability*, Volumes 1 and 2, (Translators: A. Machi and A. Smith), London, Chichester: Wiley

[25] De Morgan, A. (1926), *Formal Logic*, (Editor: A.E. Taylor), London: Open Court

[26] Dennis III, S.Y. (1991), "On the Hyper-Dirichlet Type 1 and Hyper-Liouville Distributions", *Communications in Statistics - Theory and Methods* **20**, 4069–4081

[27] Dickey, J.M. (1983), "Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses", *Journal of the American Statistical Association* **78**, 628–637

[28] Fienberg, S.E. and Holland, P.W. (1973), "Simultaneous Estimation of Multinomial Cell Probabilities", *Journal of the American Statistical Society* **68**, 683–691

[29] Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, 3rd edition, New York: Wiley

[30] Fletcher, R. (1987), *Practical Methods of Optimization*, 2nd edition, Chichester: Wiley

[31] Freedman, D.A. and Purves, R.A. (1969), "Bayes' Method for Bookies", *The Annals of Mathematical Statistics* **40**, 1177–1186

[32] Geisser, S. (1971), "The Inferential Use of Predictive Distributions", In *Foundations of Statistical Inference*, (Editors: V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 456–469

[33] Geisser, S. (1982), "Aspects of the Predictive and Estimative Approaches in the Determination of Probabilities", *Biometrics Supplement: Current Topics in Biostatistics and Epidemiology*, 75–85

[34] Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association* **85**, 398–409

[35] Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741

[36] Gilio, A. (1992), "$C_0$-Coherence and Extensions of Conditional Probabilities", In *Bayesian Statistics* **4**, (Editors: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Clarendon Press, 633–640

[37] Gong, G. and Samaniego, F.J. (1981), "Pseudo Maximum Likelihood Estimation: Theory and Applications", *The Annals of Statistics* **9**, 861–869

[38] Good, I.J. (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge, Massachusetts: The Massachusetts Institute of Technology Press

[39] Good, I.J. (1966), "How to Estimate Probabilities", *Journal of the Institute of Mathematics and its Applications* **2**, 364–383

[40] Good, I.J. (1967), "A Bayesian Significance Test for Multinomial Distributions", *Journal of the Royal Statistical Society*, Series B **29**, 399–418

[41] Good, I.J. (1976), "On the Application of Symmetric Dirichlet Distributions and their Mixtures to Contingency Tables", *The Annals of Statistics* **4**, 1159–1189

[42] Good, I.J. (1980), "On the Application of Symmetric Dirichlet Distributions and their Mixtures to Contingency Tables, Part II", *The Annals of Statistics* **8**, 1198–1218

[43] Good, I.J. (1980), "Some History of the Hierarchical Bayesian Methodology", In *Bayesian Statistics*, (Editors: J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith), Valencia: University Press, 489–504

[44] Good, I.J. and Crook, J.F. (1974), "The Bayes/Non-Bayes Compromise and the Multinomial Distribution", *Journal of the American Statistical Society* **69**, 711–720

[45] Goodhardt, G.J., Ehrenberg, A.S.C. and Chatfield, C. (1984), "The Dirichlet: A Comprehensive Model of Buying Behaviour", *Journal of the Royal Satistical Society*, Series A, **147**, 621–643

[46] Griffiths, D.A. (1973), "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease", *Biometrics* **29**, 637–648

[47] Haseman, J.K. and Kupper, L.L. (1979), "Analysis of Dichotomous Response Data from Certain Toxicological Experiments", *Biometrics* **35**, 281–293

[48] Hastings, W.K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and their Applications", *Biometrika* **57**, 97–109

[49] Heath, D. and Sudderth, W. (1978), "On Finitely Additive Priors, Coherence, and Extended Admissibility", *The Annals of Statistics* **6**, 333–345

[50] Hoadley, B. (1969), "The Compound Multinomial Distribution and Bayesian Analysis of Categorical Data from Finite Populations", *Journal of the American Statistical Society* **64**, 216–229

[51] Janardan, K.G. and Patil, G.P. (1970), "On the Multivariate Polya Distribution: A Model of Contagion for Data with Multiple Counts", In *Random Counts in Scientific Work*, Volume 3 (Editor: G.P. Patil), University Park, Pennsylvania: Pennsylvania State University Press, 143–161

[52] Jeffreys, H. (1961), *Theory of Probability*, 3rd edition, Oxford: Clarendon Press

[53] Keynes, J.M. (1921), *Treatise on Probability*, London: MacMillan

[54] Kleinman, J.C. (1973), "Proportions with Extraneous Variance: Single and Independent Samples", *Journal of the American Statistical Association* **68**, 46–54

[55] Koehler, K.J. and Wilson, J.R. (1986), "Chi-Square Tests for Comparing Vectors of Proportions for Several Cluster Samples", *Communications in Statistics - Theory and Methods* **15**, 2977–2990

[56] Lad, F. (1996), *Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction*, New York: Wiley

[57] Lad, F., Deely, J. and Piesse, A. (1993), *Using the Fundamental Theorem of Prevision to Identify Coherency Conditions for Finite Exchangeable Inference*, Research Report No. 95, Department of Mathematics and Statistics, University of Canterbury, New Zealand

[58] Lad, F., Deely, J. and Piesse, A. (1995), "Coherency Conditions for Finite Exchangeable Inference", *Journal of Italian Statistical Society* **4**, 195–213

[59] Lane, D.A. (1981), "Coherence and Prediction", *Bulletin of the Indian Standards Institution*, Proceedings of the 43rd Session, Book 1, 81–96

[60] Lane, D.A. and Sudderth, W.D. (1984), "Coherent Predictive Inference", *Sankhyā* **46**, Series A, 166–185

[61] Lane, D.A. and Sudderth, W.D. (1985), "Coherent Predictions are Strategic", *The Annals of Statistics* **13**, 1244–1248

[62] Laplace, P.S. (1812), *Théorie Analytique des Probabilités*, Paris: Courcier

[63] Leckenby, J.D. and Kishi, S. (1984), "The Dirichlet Multinomial Distribution as a Magazine Exposure Model", *Journal of Marketing Research* **21**, 100–106

[64] Lee, J.C. and Sabavala, D.J. (1987), "Bayesian Estimation and Prediction for the Beta-Binomial Model", *Journal of Business and Economic Statistics* **5**, 357–368

[65] Lenk, P.J. (1992), "Hierarchical Bayes Forecasts of Multinomial Dirichlet Data Applied to Coupon Redemptions", *Journal of Forecasting* **11**, 603–619

[66] Leonard, T. (1977), "Bayesian Simultaneous Estimation for Several Multinomial Distributions", *Communications in Statistics - Theory and Methods* **6**, 619–630

[67] Leonard, T. (1977), "A Bayesian Approach to Some Multinomial Estimation and Pretesting Problems", *Journal of the American Statistical Association* **72**, 869–874

[68] Levin, B. and Reeds, J. (1977), "Compound Multinomial Likelihood Functions are Unimodal: Proof of a Conjecture of I.J. Good", *The Annals of Statistics* **5**, 79–87

[69] Lidstone, G.J. (1920), "Note on the General Case of the Bayes-Laplace Formula for Inductive or *A Posteriori* Probabilities", *Transactions of the Faculty of Actuaries* **8**, 182–192

[70] Lindley, D.V. (1964), "The Bayesian Analysis of Contingency Tables", *The Annals of Mathematical Statistics* **35**, 1622–1643

[71] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics* **21**, 1087–1091

[72] Mosimann, J.E. (1962), "On the Compound Multinomial Distribution, the Multivariate $\beta$-distribution, and Correlations Among Proportions", *Biometrika* **49**, 65–82

[73] Paul, S.R., Liang, K.Y. and Self, S.G. (1989), "On Testing Departure from the Binomial and Multinomial Assumptions", *Biometrics* **45**, 231–236

[74] Perks, W. (1947), "Some Observations on Inverse Probability Including a New Indifference Rule", *Journal of the Institute of Actuaries* **73**, 285–312

[75] Pidd, M. (1984), *Computer Simulation in Management Science*, Chichester: Wiley

[76] Ramsey, F.P. (1931), *The Foundations of Mathematics and other Logical Essays*, London: Routledge and Kegan Paul

[77] Rao, J.N.K. and Scott, A.J. (1981), "The Analysis of Categorical Data From Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", *Journal of the American Statistical Association* **76**, 221–230

[78] Regazzini, E. (1987), "De Finetti's Coherence and Statistical Inference", *The Annals of Statistics* **15**, 845–864

[79] Rudin, W. (1976), *Principles of Mathematical Analysis*, 3rd edition, New York: McGraw-Hill

[80] Rust, R.T. and Leone, R.P. (1984), "The Mixed Media Dirichlet Multinomial Distribution: A Model for Evaluating Television-Magazine Advertising Schedules", *Journal of Marketing Research* **21**, 89–99

[81] Savage, L.J. (1972), *The Foundations of Statistics*, 2nd edition, New York: Dover

[82] Segreti, A.C. and Munson, A.E. (1981), "Estimation of the Median Lethal Dose When Response within a Litter are Correlated", *Biometrics* **37**, 153–156

[83] Smith, A.F.M. and Gelfand, A.E. (1992), "Bayesian Statistics Without Tears: A Sampling-Resampling Perspective", *The American Statistician* **46**, 84–88

[84] Stone, M. (1974), "Cross-Validation and Multinomial Prediction", *Biometrika* **61**, 509–515

[85] Trybula, S. (1958), "Some Problems of Simultaneous Minimax Estimation", *The Annals of Mathematical Statistics* **29**, 245–253

[86] Unkelbach, H.D. (1980), "The Statistical Analysis of the Differential Blood Count", *Biometrical Journal* **22**, 545–552

[87] Venn, J. (1888), *The Logic of Chance*, 3rd edition, London: MacMillan

[88] *The VNR Concise Encyclopedia of Mathematics* (1977), (Editors: W. Gellert, H. Küstner, M. Hellwich, H. Kästner), New York: Van Nostrand Reinhold

[89] Wilks, S.S. (1962), *Mathematical Statistics*, New York: Wiley

[90] Williams, D.A. (1975), "The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity", *Biometrics* **31**, 949–952

[91] Wilson, J.R. (1986), "Approximate Distribution and Test of Fit for the Clustering Effect in the Dirichlet Multinomial Model", *Communications in Statistics - Theory and Methods* **15**, 1235–1249

# Appendix A

# Programs

## A.1 SetUpMos

```
# Reads file Parameters which is assumed to assign the values of
# N+1 >= 3 (Nplus1), r >=1 and K+1 >= 2 (Kplus1). Finds the set of
# valid conditioning histograms using Mosimann's formulae and derives
# the associated system of coherency induced equations, along with the
# condition that all of the variables sum to 1.

read Parameters;

N := Nplus1 - 1:
K := Kplus1 - 1:

GenerateTypes := proc(r, Kplus1)
# Creates an array called Types of all possible outcomes for a group,
# each written as a list. Creates a global variable called dimtype,
# which is the number of possible outcomes.
global dimtype, K, Types;
local Lst, d, Lstperm, Typs, count, x, j;
    Lst := [r - trunc(d/Kplus1) $ d = 0 .. (r+1)*Kplus1 - 1];
    Lstperm := combinat[permute](Lst, Kplus1);
```

```
    dimtype := (r+K)!/(r!*K!);
    Typs := array(1 .. dimtype);
    count := 0;
    for x from 1 to nops(Lstperm) do
        if sum(Lstperm[x][j], j = 1 .. Kplus1) = r then
            count := count + 1;
            Typs[count] := Lstperm[x]
        fi
    od;
    Types := convert(Typs, list);
end:

GenerateQs := proc(colnumber, colsleft, typenumber)
# Creates a table called Qlist containing outcomes of the N+1 groups
# for which some subset of N of them form a valid conditioning
# histogram using Mosimann's formulae.  Each outcome is represented
# by a list of terms [v,w] meaning that v of the N+1 groups produced
# a type w.  Creates a global variable called qcount, which is the
# dimension of Qlist.  Creates a table called Qbox with entries of
# the form [v,w] as described above (Qbox is only global due to
# recursive procedure calls and is not otherwise used).
global Qbox, Qlist, qcount, dimtype;
local Tempqbox, d, c;
    if colsleft = 0 then
        Tempqbox := [Qbox[d] $ d  = 1 .. colnumber - 1];
        if TestQ(Tempqbox) > 0 then
            Qlist[qcount] := Tempqbox;
            qcount := qcount + 1
        fi
    else
        if typenumber = dimtype then
            Qbox[colnumber] := [colsleft, typenumber];
            GenerateQs(colnumber + 1, 0, 0)
```

```
        else
            for c from colsleft by -1 to 0 do
                if c > 0 then
                    Qbox[colnumber] := [c, typenumber];
                    GenerateQs(colnumber + 1, colsleft - c,
                     typenumber + 1)
                else
                    GenerateQs(colnumber, colsleft, typenumber + 1)
                fi
            od
        fi
    fi
end:

TestQ := proc(Tempqbox)
# Returns 0 if Tempqbox is an outcome of the N+1 groups for which no
# subset of N of them form a valid conditioning histogram using
# Mosimann's formulae and a positive number otherwise.  Creates a
# table called Storehists of all valid conditioning histograms using
# Mosimann's formulae, along with, for each histogram, a list of
# indices identifying which variables from Qlist are associated with
# that histogram, the corresponding value of Mosimann's estimate of
# tau and a list of the average number of items observed in each
# category.  Creates a global variable called histcount, which is the
# dimension of Storehists.
global K, Types, N, Kplus1, r, histcount, Storehists, dimtype,
 qcount;
local z, Temp, Hist, Chat, tauht, tauhat, Wm, Wdm, j, tally, x,
 Ybar, k, Test, flag, h, d, testQ;
    for z from 1 to nops(Tempqbox) do
        Temp := [Tempqbox[z][1] - 1, Tempqbox[z][2]];
        if Temp[1] > 0 then
            Hist := subsop(z = Temp, Tempqbox)
```

```
else
    Hist := subsop(z = NULL, Tempqbox)
fi;
Chat := 'Chat';
tauht := 'tauht';
tauhat := 'tauhat';
Wm := array(symmetric, 1 .. K, 1 .. K);
Wdm := array(symmetric, 1 .. K, 1 .. K);
for j from 1 to K do
    tally := 0;
    for x from 1 to nops(Hist) do
        tally := tally + (Hist[x][1])*(Types[Hist[x][2]][j])
    od;
    Ybar[j] := tally/N;
    if Ybar[j] = 0 then
        Chat := 0;
        break
    fi
od;
j := 'j';
Ybar[Kplus1] := r - sum(Ybar[j], j = 1 .. K);
if Chat <> 0 then
    for j from 1 to K do
        for k from j to K do
            if k = j then
                Wm[j,k] := Ybar[j]*(r - Ybar[j])/r;
                x := 'x';
                Wdm[j,k] := (1/(N-1))*sum((Hist[x][1])*
                 (Types[Hist[x][2]][j] - Ybar[j])^2, x = 1 ..
                 nops(Hist))
            else
                Wm[j,k] := -Ybar[j]*Ybar[k]/r;
                x := 'x';
```

```
                    Wdm[j,k] := (1/(N-1))*sum((Hist[x][1])*
                     ((Types[Hist[x][2]][j] - Ybar[j])*
                     (Types[Hist[x][2]][k] - Ybar[k])), x = 1 ..
                     nops(Hist))
                fi
            od
        od;
        with(linalg);
        if det(Wm) = 0 then
            Chat := 0
        else
            Chat := abs((det(Wdm)/det(Wm))^(1/K));
        fi
    fi;
    if Chat = 1 then
        Test[z] := 0
    else
        tauht := (r - Chat)/(Chat - 1);
        tauhat := simplify(tauht);
        if evalf(tauhat) > 0 then
            Test[z] := 1
        else
            Test[z] := 0
        fi
    fi;
    if Test[z] = 1 then
        flag := 0;
        for h from 1 to histcount do
            if nops(Storehists[h][2]) < dimtype then
                if Hist = Storehists[h][1] then
                    Temp := [Storehists[h][2][d] $ d = 1 ..
                     nops(Storehists[h][2]), qcount];
                    Storehists[h] := subsop(2 = Temp,
```

```
                        Storehists[h]);
                    flag := 1;
                    break
                fi
            fi
        od;
        if flag = 0 then
            j := 'j';
            Storehists[histcount] := [Hist, [qcount], tauhat,
             [Ybar[j] $ j = 1 .. Kplus1]];
            histcount := histcount + 1
        fi
    fi
od;
testQ := sum(Test[d], d = 1 .. nops(Tempqbox));
testQ
end:

GenerateEquations := proc()
# Creates a table of tables called Probs which contains, for each
# valid conditioning histogram, the predictive probabilities of the
# last group's outcomes being of each possible type.  Creates all of
# the coherency induced equations, including the last condition that
# all of the variables sum to 1.  These equations are called
# e1, e2, ...
global histcount, Storehists, Kplus1, r, dimtype, Types, Probs,
 Nplus1, dimq;
local h, tauhat, j, Ybar, Alphahat, t, Part1prob, result, ans,
 Part2prob, f, x, s, u;
    for h from 1 to histcount do
        tauhat := Storehists[h][3];
        for j from 1 to Kplus1 do
            Ybar[j] := Storehists[h][4][j];
```

```
    Alphahat[j] := tauhat*Ybar[j]/r;
od;
for t from 1 to dimtype do
    j := 'j';
    Part1prob[h][t] := simplify((r!/product(Types[t][j]!,
     j = 1 .. Kplus1))/Prod(tauhat, r));
    result := 1;
    for j from 1 to Kplus1 do
        ans := Prod(Alphahat[j], Types[t][j]);
        result := result*ans
    od;
    Part2prob[h][t] := simplify(result);
    Probs[h][t] := Part1prob[h][t]*Part2prob[h][t];
od;
t := 't';
for t from 1 to dimtype do
    f[h][t] := 0
od;
t := 't';
for x from 1 to nops(Storehists[h][1]) do
    t := Storehists[h][1][x][2];
    f[h][t] := (Storehists[h][1][x][1] + 1)/Nplus1
od;
t := 't';
for t from 1 to dimtype do
    if f[h][t] > 0 then
        next
    else
        f[h][t] := 1/Nplus1
    fi
od;
t := 't';
for t from 1 to (dimtype - 1) do
```

```
            e.((dimtype-1)*(h-1) + t) := simplify(sum(f[h][s]*
             (-Probs[h][t])*q.(Storehists[h][2][s]), s = 1 ..
             dimtype) + f[h][t]*q.(Storehists[h][2][t]) = 0);
        od
    od;
    e.((dimtype-1)*histcount + 1) := sum('q.(u)', u = 1 .. dimq) = 1
end:

Prod := proc(startpt, maxint)
# Returns the value of Gamma(startpt + maxint)/Gamma(startpt).
global answer;
local d;
    if maxint = 0 then
        1
    else
        answer := 1;
        for d from 0 to (maxint - 1) do
            answer := answer*(startpt + d)
        od
    fi
end:

GenerateTypes(r, Kplus1):

qcount := 1:

histcount := 1:

Storehists[1] := [[], [0]]:

GenerateQs(1, Nplus1, 1):

Qbox := 'Qbox':
```

```
dimtype := (r+K)!/(r!*K!);

histcount := histcount - 1;

qcount := qcount - 1;

dimq := (Nplus1 + dimtype-1)!/(Nplus1!*(dimtype-1)!);

eqncount := (dimtype-1)*histcount + 1;

GenerateEquations():


# Solution := solve({e.(1 .. eqncount - 1)}, {q.(1 .. qcount)});
```

## A.2 SetUpBrier

```
# Reads file Parameters which is assumed to assign the values of
# N+1 >= 3 (Nplus1), r >=1 and K+1 >= 2 (Kplus1).  Finds the set of
# valid conditioning histograms using Brier's formulae and derives
# the associated system of coherency induced equations, along with
# the condition that all of the variables sum to 1.

read Parameters;

N := Nplus1 - 1:
K := Kplus1 - 1:

GenerateTypes := proc(r, Kplus1)
# Creates an array called Types of all possible outcomes for a group,
```

```
# each written as a list.  Creates a global variable called dimtype,
# which is the number of possible outcomes.
global dimtype, K, Types;
local Lst, d, Lstperm, Typs, count, x, j;
    Lst := [r - trunc(d/Kplus1) $ d = 0 .. (r+1)*Kplus1 - 1];
    Lstperm := combinat[permute](Lst, Kplus1);
    dimtype := (r+K)!/(r!*K!);
    Typs := array(1 .. dimtype);
    count := 0;
    for x from 1 to nops(Lstperm) do
        if sum(Lstperm[x][j], j = 1 .. Kplus1) = r then
            count := count + 1;
            Typs[count] := Lstperm[x]
        fi
    od;
    Types := convert(Typs, list);
end:

GenerateQs := proc(colnumber, colsleft, typenumber)
# Creates a table called Qlist containing outcomes of the N+1 groups
# for which some subset of N of them form a valid conditioning
# histogram using Brier's formulae.  Each outcome is represented by a
# list of terms [v,w] meaning that v of the N+1 groups produced a
# type w.  Creates a global variable called qcount, which is the
# dimension of Qlist.  Creates a table called Qbox with entries of
# the form [v,w] as described above (Qbox is only global due to
# recursive procedure calls and is not otherwise used).
global Qbox, Qlist, qcount, dimtype;
local Tempqbox, d, c;
    if colsleft = 0 then
        Tempqbox := [Qbox[d] $ d  = 1 .. colnumber - 1];
        if TestQ(Tempqbox) > 0 then
            Qlist[qcount] := Tempqbox;
```

```
            qcount := qcount + 1
        fi
    else
        if typenumber = dimtype then
            Qbox[colnumber] := [colsleft, typenumber];
            GenerateQs(colnumber + 1, 0, 0)
        else
            for c from colsleft by -1 to 0 do
                if c > 0 then
                    Qbox[colnumber] := [c, typenumber];
                    GenerateQs(colnumber + 1, colsleft - c,
                     typenumber + 1)
                else
                    GenerateQs(colnumber, colsleft, typenumber + 1)
                fi
            od
        fi
    fi
end:

TestQ := proc(Tempqbox)
# Returns 0 if Tempqbox is an outcome of the N+1 groups for which no
# subset of N of them form a valid conditioning histogram using
# Brier's formulae and a positive number otherwise.  Creates a table
# called Storehists of all valid conditioning histograms using
# Brier's formulae, along with, for each histogram, a list of indices
# identifying which variables from Qlist are associated with that
# histogram, the corresponding value of Brier's estimate of tau and a
# list of the average number of items observed in each category.
# Creates a global variable called histcount, which is the dimension
# of Storehists.
global K, Types, N, Kplus1, r, histcount, Storehists, dimtype,
 qcount;
```

```
local z, Temp, Hist, Chat, tauhat, j, tally, x, Ybar, Test, flag, h,
 d, testQ;
    for z from 1 to nops(Tempqbox) do
        Temp := [Tempqbox[z][1] - 1, Tempqbox[z][2]];
        if Temp[1] > 0 then
            Hist := subsop(z = Temp, Tempqbox)
        else
            Hist := subsop(z = NULL, Tempqbox)
        fi;
        Chat := 'Chat';
        tauhat := 'tauhat';
        for j from 1 to K do
            tally := 0;
            for x from 1 to nops(Hist) do
                tally := tally + (Hist[x][1])*(Types[Hist[x][2]][j])
            od;
            Ybar[j] := tally/N;
            if Ybar[j] = 0 then
                Chat := 0;
                break
            fi
        od;
        j := 'j';
        Ybar[Kplus1] := r - sum(Ybar[j], j = 1 .. K);
        if Ybar[Kplus1] = 0 then
            Chat := 0
        fi;
        if Chat <> 0 then
            x := 'x';
            Chat := (1/((N-1)*K))*sum(sum((Hist[x][1])*
             ((Types[Hist[x][2]][j] - Ybar[j])^2/Ybar[j]),
             x = 1 .. nops(Hist)), j = 1 .. Kplus1)
        fi;
```

```
        if Chat = 1 then
            Test[z] := 0
        else
            tauhat := (r - Chat)/(Chat - 1);
            if tauhat > 0 then
                Test[z] := 1
            else
                Test[z] := 0
            fi
        fi;
        if Test[z] = 1 then
            flag := 0;
            for h from 1 to histcount do
                if nops(Storehists[h][2]) < dimtype then
                    if Hist = Storehists[h][1] then
                        Temp := [Storehists[h][2][d] $ d = 1 ..
                         nops(Storehists[h][2]), qcount];
                        Storehists[h] := subsop(2 = Temp,
                         Storehists[h]);
                        flag := 1;
                        break
                    fi
                fi
            od;
            if flag = 0 then
                j := 'j';
                Storehists[histcount] := [Hist, [qcount], tauhat,
                 [Ybar[j] $ j = 1 .. Kplus1]];
                histcount := histcount + 1
            fi
        fi
    od;
    testQ := sum(Test[d], d = 1 .. nops(Tempqbox));
```

```
    testQ
end:

GenerateEquations := proc()
# Creates a table of tables called Probs which contains, for each
# valid conditioning histogram, the predictive probabilities of the
# last group's outcomes being of each possible type.  Creates all of
# the coherency induced equations, including the last condition that
# all of the variables sum to 1.  These equations are called
# e1, e2, ....
global histcount, Storehists, Kplus1, r, dimtype, Probs, Types,
 Nplus1, dimq;
local h, tauhat, j, Ybar, Alphahat, t, d, f, x, s, u;
    for h from 1 to histcount do
        tauhat := Storehists[h][3];
        for j from 1 to Kplus1 do
            Ybar[j] := Storehists[h][4][j];
            Alphahat[j] := tauhat*Ybar[j]/r
        od;
        for t from 1 to dimtype do
            j := 'j';
            Probs[h][t] := (r!/product(Types[t][j]!, j = 1 ..
             Kplus1))*(product(Prod(t, j, Alphahat), j = 1 ..
             Kplus1)/product(tauhat + d, d = 0 .. r - 1))
        od;
        t := 't';
        for t from 1 to dimtype do
            f[h][t] := 0
        od;
        t := 't';
        for x from 1 to nops(Storehists[h][1]) do
            t := Storehists[h][1][x][2];
            f[h][t] := (Storehists[h][1][x][1] + 1)/Nplus1
```

```
        od;
        t := 't';
        for t from 1 to dimtype do
            if f[h][t] > 0 then
                next
            else
                f[h][t] := 1/Nplus1
            fi
        od;
        t := 't';
        for t from 1 to (dimtype - 1) do
            e.((dimtype-1)*(h-1) + t) := sum(f[h][s]*(-Probs[h][t])*
             q.(Storehists[h][2][s]), s = 1 .. dimtype) +
             f[h][t]*q.(Storehists[h][2][t]) = 0;
        od
    od;
    e.((dimtype-1)*histcount + 1) := sum('q.(u)', u = 1 .. dimq) = 1
end:

Prod := proc(t, j, Alphahat)
# Returns the value of Gamma(Alphahat[j] + Types[t][j])/
# Gamma(Alphahat[j]).
global Types;
local d;
    if Types[t][j] = 0 then
        1
    else
        product(Alphahat[j] + d, d = 0 .. Types[t][j] - 1)
    fi
end:

GenerateTypes(r, Kplus1):
```

```
qcount := 1:

histcount := 1:

Storehists[1] := [[], [0]]:

GenerateQs(1, Nplus1, 1):

Qbox := 'Qbox':

dimtype := (r+K)!/(r!*K!);

histcount := histcount - 1;

qcount := qcount - 1;

dimq := (Nplus1 + dimtype-1)!/(Nplus1!*(dimtype-1)!);

eqncount := (dimtype-1)*histcount + 1;

GenerateEquations():


# Solution := solve({e.(1 .. eqncount - 1)}, {q.(1 .. qcount)});
```

## A.3 TestLinked

```
# Assumes that another program, such as SetUpMos, has created the
# system of equations associated with histograms from some set and
# tests whether or not this system (or set) is linked.
```

```
Links := proc()
# Creates a table called Tag containing, for each variable in the
# equations, a list of the numbers of the histograms whose equation
# blocks it appears in.
global qcount, Tag, histcount, dimtype, Storehists;
local q, h, t;
    for q from 1 to qcount do
        Tag[q] := []
    od;
    for h from 1 to histcount do
        for t from 1 to dimtype do
            Tag[Storehists[h][2][t]] :=
             [op(Tag[Storehists[h][2][t]]), h]
        od
    od
end:

AdjacentHists := proc()
# Creates a table called Adjhists containing, for each conditioning
# histogram, a list of the other histograms that it is linked to.
global histcount, Adjhists, qcount, Tag, dimtype;
local h, q, x, d, position;
    for h from 1 to histcount do
        Adjhists[h] := []
    od;
    for q from 1 to qcount do
        if nops(Tag[q]) > 1 then
            for x from 1 to nops(Tag[q]) do
                Adjhists[Tag[q][x]] := [op(Adjhists[Tag[q][x]]),
                 Tag[q][d] $ d = 1 .. nops(Tag[q])]
            od
        fi
    od;
```

```
    h := 'h';
    for h from 1 to histcount do
        Adjhists[h] := sort(Adjhists[h]);
        for d from 1 to dimtype while member(h, Adjhists[h]) do
            member(h, Adjhists[h], 'position');
            Adjhists[h] := subsop(position = NULL, Adjhists[h])
        od
    od
end:

TestEqnsLinked := proc()
# Returns 1 if the set of conditioning histograms (or associated
# equations) is linked and 0 otherwise.
global histcount, Adjhists;
local Histlink, Hset, d, h, x, linked;
    Histlink := {1};
    Hset := {};
    for d from 1 to histcount while nops(Histlink) < histcount do
        h := min(op(Histlink minus Hset));
        if h = infinity then
            break
        fi;
        Hset := Hset union {h};
        for x from 1 to nops(Adjhists[h]) do
            Histlink := Histlink union {Adjhists[h][x]} union
             convert(Adjhists[Adjhists[h][x]], set)
        od
    od;
    if nops(Histlink) = histcount then
        linked := 1
    else
        linked := 0
    fi
```

```
end:

Links():

AdjacentHists():

TestEqnsLinked();
```

## A.4 SetUpNoTau

```
# Reads file Parameters which is assumed to assign the values of
# N+1 >= 3 (Nplus1), r >=1 and K+1 >= 2 (Kplus1).  Finds the set of
# positive conditioning histograms and derives the associated system
# of coherency induced equations, along with the condition that all
# of the variables sum to 1.

read Parameters;

N := Nplus1 - 1:
K := Kplus1 - 1:

GenerateTypes := proc(r, Kplus1)
# Creates an array called Types of all possible outcomes for a group,
# each written as a list.  Creates a global variable called dimtype,
# which is the number of possible outcomes.
global dimtype, K, Types;
local Lst, d, Lstperm, Typs, count, x, j;
    Lst := [r - trunc(d/Kplus1) $ d = 0 .. (r+1)*Kplus1 - 1];
    Lstperm := combinat[permute](Lst, Kplus1);
    dimtype := (r+K)!/(r!*K!);
    Typs := array(1 .. dimtype);
```

```
    count := 0;
    for x from 1 to nops(Lstperm) do
        if sum(Lstperm[x][j], j = 1 .. Kplus1) = r then
            count := count + 1;
            Typs[count] := Lstperm[x]
        fi
    od;
    Types := convert(Typs, list);
end:

GenerateQs := proc(colnumber, colsleft, typenumber)
# Creates a table called Qlist containing outcomes of the N+1 groups
# for which some subset of N of them form a positive histogram.  Each
# outcome is represented by a list of terms [v,w] meaning that v of
# the N+1 groups produced a type w.  Creates a global variable called
# qcount, which is the dimension of Qlist.  Creates a table called
# Qbox with entries of the form [v,w] as described above (Qbox is
# only global due to recursive procedure calls and is not otherwise
# used).
global Qbox, Qlist, qcount, dimtype;
local Tempqbox, d, c;
    if colsleft = 0 then
        Tempqbox := [Qbox[d] $ d  = 1 .. colnumber - 1];
        if TestQ(Tempqbox) > 0 then
            Qlist[qcount] := Tempqbox;
            qcount := qcount + 1
        fi
    else
        if typenumber = dimtype then
            Qbox[colnumber] := [colsleft, typenumber];
            GenerateQs(colnumber + 1, 0, 0)
        else
            for c from colsleft by -1 to 0 do
```

```
                if c > 0 then
                    Qbox[colnumber] := [c, typenumber];
                    GenerateQs(colnumber + 1, colsleft - c,
                     typenumber + 1)
                else
                    GenerateQs(colnumber, colsleft, typenumber + 1)
                fi
            od
        fi
    fi
end:

TestQ := proc(Tempqbox)
# Returns 0 if Tempqbox is an outcome of the N+1 groups for which no
# subset of N of them form a positive histogram and a positive number
# otherwise.  Creates a table called Simhists whose entries are the
# sets of positive histograms that are equivalent under a permutation
# of category ordering.  Creates a table called Storehists of all
# positive histograms, along with, for each histogram, a list of
# indices identifying which variables from Qlist are associated with
# that histogram, an entry of the form T.v where v identifies which
# set from Simhists the histogram belongs to, and a list of the
# average number of items observed in each category.  Creates global
# variables called histno and histcount, which are the dimension of
# Simhists and Storehists, respectively.
global K, Types, N, Kplus1, r, histno, Simhists, dimtype, histcount,
 Storehists, qcount;
local z, Temp, Hist, tauhat, j, tally, x, Ybar, Test, flag, g, Rows,
 d, Rowperms, w, i, t, Grptypes, Grouptypes, Smhists, typecount, h,
 testQ;
    for z from 1 to nops(Tempqbox) do
        Temp := [Tempqbox[z][1] - 1, Tempqbox[z][2]];
        if Temp[1] > 0 then
```

```
    Hist := subsop(z = Temp, Tempqbox)
else
    Hist := subsop(z = NULL, Tempqbox)
fi;
tauhat := 'tauhat';
for j from 1 to K do
    tally := 0;
    for x from 1 to nops(Hist) do
        tally := tally + (Hist[x][1])*(Types[Hist[x][2]][j])
    od;
    Ybar[j] := tally/N;
    if Ybar[j] = 0 then
        tauhat := 0;
        break
    fi
od;
j := 'j';
Ybar[Kplus1] := r - sum(Ybar[j], j = 1 .. K);
if Ybar[Kplus1] = 0 then
    tauhat := 0
fi;
if tauhat = 0 then
    Test[z] := 0
else
    flag := 0;
    for g from 1 to (histno - 1) while flag = 0 do
      if member(Hist, Simhists[g]) then
            tauhat := T.g;
            Test[z] := 1;
            flag := 1;
            break
        fi
    od;
```

```
if flag = 0 then
    x := 'x';
    for j from 1 to Kplus1 do
        Rows[j] := [];
        for x from 1 to nops(Hist) do
            Rows[j] := [op(Rows[j]),
             (Types[Hist[x][2]][j] $ d = 1 .. Hist[x][1])]
        od
    od;
    Rowperms := combinat[permute](convert(Rows, list));
    j := 'j';
    for w from 1 to nops(Rowperms) do
        for i from 1 to N do
            for t from 1 to dimtype do
                if [Rowperms[w][j][i] $ j = 1 .. Kplus1] =
                 Types[t] then
                    Grptypes[i] := t;
                    break
                fi
            od
        od;
        Grouptypes := sort(convert(Grptypes, list));
        typecount := 0;
        Smhists[w] := [];
        for i from 1 to (N - 1) do
            typecount := typecount + 1;
            if Grouptypes[i+1] <> Grouptypes[i] then
                Smhists[w] := [op(Smhists[w]),
                 [typecount, Grouptypes[i]]];
                typecount := 0
            fi
        od;
        typecount := typecount + 1;
```

```
                Smhists[w] := [op(Smhists[w]), [typecount,
                 Grouptypes[N]]]
            od;
            Simhists[histno] := convert(Smhists, set);
            tauhat := T.histno;
            Test[z] := 1;
            histno := histno + 1
        fi
    fi;
    flag := 'flag';
    if Test[z] = 1 then
        flag := 0;
        for h from 1 to histcount do
            if nops(Storehists[h][2]) < dimtype then
                if Hist = Storehists[h][1] then
                    Temp := [Storehists[h][2][d] $ d = 1 ..
                     nops(Storehists[h][2]), qcount];
                    Storehists[h] := subsop(2 = Temp,
                     Storehists[h]);
                    flag := 1;
                    break
                fi
            fi
        od;
        if flag = 0 then
            j := 'j';
            Storehists[histcount] := [Hist, [qcount], tauhat,
             [Ybar[j] $ j = 1 .. Kplus1]];
            histcount := histcount + 1
        fi
    fi
od;
testQ := sum(Test[d], d = 1 .. nops(Tempqbox));
```

```
    testQ
end:

GenerateEquations := proc()
# Creates a table of tables called Probs containing, for each
# positive conditioning histogram, the predictive probabilities of
# the last group's outcomes being of each possible type.  Creates all
# of the coherency induced equations, including the last condition
# that all of the variables sum to 1.  These equations are called
# e1, e2, ...
global histcount, Storehists, Kplus1, r, dimtype, Types, Probs,
 Nplus1, dimq;
local h, tauhat, j, Ybar, Alphahat, t, Part1prob, result, ans,
 Part2prob, f, x, s, u;
    for h from 1 to histcount do
        tauhat := Storehists[h][3];
        for j from 1 to Kplus1 do
            Ybar[j] := Storehists[h][4][j];
            Alphahat[j] := tauhat*Ybar[j]/r;
        od;
        for t from 1 to dimtype do
            j := 'j';
            Part1prob[h][t] := simplify((r!/product(Types[t][j]!,
             j = 1 .. Kplus1))/Prod(tauhat, r));
            result := 1;
            for j from 1 to Kplus1 do
                ans := Prod(Alphahat[j], Types[t][j]);
                result := result*ans
            od;
            Part2prob[h][t] := simplify(result);
            Probs[h][t] := Part1prob[h][t]*Part2prob[h][t];
        od;
        t := 't';
```

```
        for t from 1 to dimtype do
            f[h][t] := 0
        od;
        t := 't';
        for x from 1 to nops(Storehists[h][1]) do
            t := Storehists[h][1][x][2];
            f[h][t] := (Storehists[h][1][x][1] + 1)/Nplus1
        od;
        t := 't';
        for t from 1 to dimtype do
            if f[h][t] > 0 then
                next
            else
                f[h][t] := 1/Nplus1
            fi
        od;
        t := 't';
        for t from 1 to (dimtype - 1) do
            e.((dimtype-1)*(h-1) + t) := sum(f[h][s]*(-Probs[h][t])*
             q.(Storehists[h][2][s]), s = 1 .. dimtype) +
             f[h][t]*q.(Storehists[h][2][t]) = 0;
        od
    od;
    e.((dimtype-1)*histcount + 1) := sum('q.(u)', u = 1 .. dimq) = 1
end:

Prod := proc(startpt, maxint)
# Returns the value of Gamma(startpt + maxint)/Gamma(startpt).
global answer;
local d;
    if maxint = 0 then
        1
    else
```

```
        answer := 1;
        for d from 0 to (maxint - 1) do
            answer := answer*(startpt + d)
        od
    fi
end:

GenerateTypes(r, Kplus1):

qcount := 1:

histno := 1:

Simhists[1] := {}:

histcount := 1:

Storehists[1] := [[], [0]]:

GenerateQs(1, Nplus1, 1):

Qbox := 'Qbox':

dimtype := (r+K)!/(r!*K!);

histno := histno - 1;

histcount := histcount - 1;

qcount := qcount - 1;

dimq := (Nplus1 + dimtype-1)!/(Nplus1!*(dimtype-1)!);
```

```
eqncount := (dimtype-1)*histcount + 1;

GenerateEquations():


# for i from 1 to qcount do
#     f.i := q.i <> 0
# od:
# for i from 1 to histno do
#     g.i := T.i <> 0
# od:
# i := 'i':

# Solution := solve({e.(1 .. eqncount - 1), f.(1 .. qcount),
#  g.(1 .. histno)}, {q.(1 .. qcount), T.(1 .. histno)});
```

## A.5 FindCycles

```
# Reads file Parameters which is assumed to assign the values of
# N+1 >= 3 (Nplus1), r >=1 and K+1 >= 2 (Kplus1).  Assumes that the
# value of size >= 3 has been assigned.  Finds the set of positive
# conditioning histograms and all of the n-cycles, where n = size,
# that are formed by these histograms.

read Parameters;

N := Nplus1 - 1:
K := Kplus1 - 1:

GenerateTypes := proc(r, Kplus1)
# Creates an array called Types of all possible outcomes for a group,
```

```
# each written as a list.  Creates a global variable called dimtype,
# which is the number of possible outcomes.
global dimtype, K, Types;
local Lst, d, Posstypes, Typs, count, x, j;
    Lst := [r - trunc(d/Kplus1) $ d = 0 .. (r+1)*Kplus1 - 1];
    Posstypes := combinat[permute](Lst, Kplus1);
    dimtype := (r+K)!/(r!*K!);
    Typs := array(1 .. dimtype);
    count := 0;
    for x from 1 to nops(Posstypes) do
        if sum(Posstypes[x][j], j = 1 .. Kplus1) = r then
            count := count + 1;
            Typs[count] := Posstypes[x]
        fi
    od;
    Types := convert(Typs, list);
end:

GenerateQs := proc(colnumber, colsleft, typenumber)
# Creates a table called Qlist containing outcomes of the N+1 groups
# for which some subset of N of them form a positive histogram.  Each
# outcome is represented by a list of terms [v,w] meaning that v of
# the N+1 groups produced a type w.  Creates a global variable called
# qcount, which is the dimension of Qlist.  Creates a table Qbox with
# entries of the form [v,w] as described above (Qbox is only global
# due to the recursive procedure calls and is not otherwise used).
global Qbox, Qlist, qcount, dimtype;
local Tempqbox, d, c;
    if colsleft = 0 then
        Tempqbox := [Qbox[d] $ d  = 1 .. colnumber - 1];
        if TestQ(Tempqbox) > 0 then
            Qlist[qcount] := Tempqbox;
            qcount := qcount + 1
```

```
        fi
    else
        if typenumber = dimtype then
            Qbox[colnumber] := [colsleft, typenumber];
            GenerateQs(colnumber + 1, 0, 0)
        else
            for c from colsleft by -1 to 0 do
                if c > 0 then
                    Qbox[colnumber] := [c, typenumber];
                    GenerateQs(colnumber + 1, colsleft - c,
                     typenumber + 1)
                else
                    GenerateQs(colnumber, colsleft, typenumber + 1)
                fi
            od
        fi
    fi
end:

TestQ := proc(Tempqbox)
# Returns 0 if Tempqbox is an outcome of the N+1 groups for which no
# subset of N of them form a positive histogram and a positive number
# otherwise.  Creates a table called Simhists whose entries are the
# sets of positive histograms that are equivalent under a permutation
# of category ordering.  Creates a table called Storehists of all
# positive histograms, along with, for each histogram, a list of
# indices identifying which variables from Qlist are associated with
# that histogram, an entry of the form T.v where v identifies which
# set from Simhists the histogram belongs to, and a list of the
# average number of items observed in each category.  Creates global
# variables called histno and histcount, which are the dimension of
# Simhists and Storehists, respectively.
global K, Types, N, Kplus1, r, histno, Simhists, dimtype, histcount,
```

```
 Storehists, qcount;
local z, Temp, Hist, tauhat, j, tally, x, Ybar, Test, flag, g, Rows,
 d, Rowperms, w, i, t, Grptypes, Grouptypes, Smhists, typecount, h,
 testQ;
    for z from 1 to nops(Tempqbox) do
        Temp := [Tempqbox[z][1] - 1, Tempqbox[z][2]];
        if Temp[1] > 0 then
            Hist := subsop(z = Temp, Tempqbox)
        else
            Hist := subsop(z = NULL, Tempqbox)
        fi;
        tauhat := 'tauhat';
        for j from 1 to K do
            tally := 0;
            for x from 1 to nops(Hist) do
                tally := tally + (Hist[x][1])*(Types[Hist[x][2]][j])
            od;
            Ybar[j] := tally/N;
            if Ybar[j] = 0 then
                tauhat := 0;
                break
            fi
        od;
        j := 'j';
        Ybar[Kplus1] := r - sum(Ybar[j], j = 1 .. K);
        if Ybar[Kplus1] = 0 then
            tauhat := 0
        fi;
        if tauhat = 0 then
            Test[z] := 0
        else
            flag := 0;
            for g from 1 to (histno - 1) while flag = 0 do
```

```
  if member(Hist, Simhists[g]) then
        tauhat := T.g;
        Test[z] := 1;
        flag := 1;
        break
    fi
od;
if flag = 0 then
    x := 'x';
    for j from 1 to Kplus1 do
        Rows[j] := [];
        for x from 1 to nops(Hist) do
            Rows[j] := [op(Rows[j]),
             (Types[Hist[x][2]][j] $ d = 1 .. Hist[x][1])]
        od
    od;
    Rowperms := combinat[permute](convert(Rows, list));
    j := 'j';
    for w from 1 to nops(Rowperms) do
        for i from 1 to N do
            for t from 1 to dimtype do
                if [Rowperms[w][j][i] $ j = 1 .. Kplus1] =
                 Types[t] then
                    Grptypes[i] := t;
                    break
                fi
            od
        od;
        Grouptypes := sort(convert(Grptypes, list));
        typecount := 0;
        Smhists[w] := [];
        for i from 1 to (N - 1) do
            typecount := typecount + 1;
```

```
                    if Grouptypes[i+1] <> Grouptypes[i] then
                        Smhists[w] := [op(Smhists[w]),
                         [typecount, Grouptypes[i]]];
                        typecount := 0
                    fi
                od;
                typecount := typecount + 1;
                Smhists[w] := [op(Smhists[w]), [typecount,
                 Grouptypes[N]]]
            od;
            Simhists[histno] := convert(Smhists, set);
            tauhat := T.histno;
            Test[z] := 1;
            histno := histno + 1
        fi
    fi;
    flag := 'flag';
    if Test[z] = 1 then
        flag := 0;
        for h from 1 to histcount do
            if nops(Storehists[h][2]) < dimtype then
                if Hist = Storehists[h][1] then
                    Temp := [Storehists[h][2][d] $ d = 1 ..
                     nops(Storehists[h][2]), qcount];
                    Storehists[h] := subsop(2 = Temp,
                     Storehists[h]);
                    flag := 1;
                    break
                fi
            fi
        od;
        if flag = 0 then
            j := 'j';
```

```
                Storehists[histcount] := [Hist, [qcount], tauhat,
                 [Ybar[j] $ j = 1 .. Kplus1]];
                histcount := histcount + 1
            fi
        fi
    od;
    testQ := sum(Test[d], d = 1 .. nops(Tempqbox));
    testQ
end:

Links := proc()
# Creates a table called Tag containing, for each variable in the
# equations, a list of the numbers of the histograms whose equation
# blocks it appears in.
global qcount, Tag, histcount, dimtype, Storehists;
local q, h, t;
    for q from 1 to qcount do
        Tag[q] := []
    od;
    for h from 1 to histcount do
        for t from 1 to dimtype do
            Tag[Storehists[h][2][t]] :=
             [op(Tag[Storehists[h][2][t]]), h]
        od
    od
end:

AdjacentHists := proc()
# Creates a table called Adjhists containing, for each conditioning
# histogram, a list of the other histograms that it is linked to.
# Creates a table called Adjhistsfol containing, for each
# conditioning histogram, a list of the histograms following it in
# order, that it is linked to.
```

```
global histcount, Adjhists, Adjhistsfol, qcount, Tag, dimtype;
local h, q, x, d, position;
    for h from 1 to histcount do
        Adjhists[h] := [];
        Adjhistsfol[h] := []
    od;
    for q from 1 to qcount do
        if nops(Tag[q]) > 1 then
            for x from 1 to nops(Tag[q]) do
                Adjhists[Tag[q][x]] := [op(Adjhists[Tag[q][x]]),
                 Tag[q][d] $ d = 1 .. nops(Tag[q])];
                Adjhistsfol[Tag[q][x]] := [op(Adjhistsfol[Tag[q][x]]),
                 Tag[q][d] $ d = x + 1 .. nops(Tag[q])]
            od
        fi
    od;
    h := 'h';
    for h from 1 to histcount do
        Adjhists[h] := sort(Adjhists[h]);
        if nops(Adjhistsfol[h]) > 1 then
            Adjhistsfol[h] := sort(Adjhistsfol[h])
        fi;
        for d from 1 to dimtype while member(h, Adjhists[h]) do
            member(h, Adjhists[h], 'position');
            Adjhists[h] := subsop(position = NULL, Adjhists[h])
        od
    od
end:

GenerateCycles := proc()
# Creates a table called Cyclesize (where size is its numerically
# assigned value) of all size-cycles.  Creates a global variable
# called cyclesizeno which is the dimension of Cyclesize.  Creates
```

```
# global variables called x, y and h (which are only global due to
# recursive procedure calls and are not otherwise used).
global size, qcount, Tag, x, y, cycle3no, h, histcount, Adjhistsfol,
 Cycle3, cycle4no, Adjhists, Cycle4;
local Samevble, q, z, Tempset, Temp, d, Chkset;
    if size = 3 then
        Samevble := {};
        for q from 1 to qcount do
            if nops(Tag[q]) > 2 then
                for x from 1 to (nops(Tag[q]) - 2) do
                    for y from (x + 1) to (nops(Tag[q]) - 1) do
                        for z from (y + 1) to nops(Tag[q]) do
                            Samevble := Samevble union {[Tag[q][x],
                             Tag[q][y], Tag[q][z]]}
                        od
                    od
                od
            fi
        od;
        x := 'x';
        z := 'z';
        cycle3no := 1;
        for h from 1 to histcount do
            if nops(Adjhistsfol[h]) > 1 then
                for x from 1 to (nops(Adjhistsfol[h]) - 1) do
                    Tempset := convert(Adjhistsfol[h], set) intersect
                     convert(Adjhistsfol[Adjhistsfol[h][x]], set);
                    for z from 1 to nops(Tempset) do
                        Temp := [h, Adjhistsfol[h][x], Tempset[z]];
                        if {Temp} intersect Samevble = {} then
                            Cycle3[cycle3no] := Temp;
                            cycle3no := cycle3no + 1
                        fi
```

```
                od
            od
        fi
    od;
    cycle3no := cycle3no - 1
elif size = 4 then
    cycle4no := 1;
    for h from 1 to (histcount - 3) do
        if nops(Adjhistsfol[h]) > 1 then
            for x from 1 to (nops(Adjhistsfol[h]) - 1) do
                for y from (x + 1) to nops(Adjhistsfol[h]) do
                    if member(Adjhistsfol[h][y],
                     Adjhistsfol[Adjhistsfol[h][x]]) = false then
                        Tempset :=
                         convert(Adjhists[Adjhistsfol[h][x]], set)
                         intersect
                         convert(Adjhists[Adjhistsfol[h][y]], set)
                         intersect {d $ d = h + 1 .. histcount};
                        for z from 1 to nops(Tempset) do
                            if member(Tempset[z], Adjhists[h]) =
                             false then
                                Cycle4[cycle4no] := [h,
                                 Adjhistsfol[h][x], Tempset[z],
                                 Adjhistsfol[h][y]];
                                cycle4no := cycle4no + 1
                            fi
                        od
                    fi
                od
            od
        fi
    od;
    cycle4no := cycle4no - 1
```

```
    else
        cycle.size.no := 1;
        for h from 1 to (histcount - size + 1) do
            if nops(Adjhistsfol[h]) > 1 then
                for x from 1 to (nops(Adjhistsfol[h]) - 1) do
                    for y from (x + 1) to nops(Adjhistsfol[h]) do
                        if member(Adjhistsfol[h][y],
                         Adjhistsfol[Adjhistsfol[h][x]]) = false then
                            Chkset := convert(Adjhists[h], set);
                            Cycles(2, size, Chkset, Adjhistsfol[h][x])
                        fi
                    od
                od
            fi
        od;
        cycle.size.no := cycle.size.no - 1
    fi
end:

Cycles := proc(place, n, checkset, testhist)
# Creates a table called Cyclesize (where size is its numerically
# assigned value) of all size-cycles.  Creates a global variable
# called cyclesizeno which is the dimension of Cyclesize.  Creates
# global variables called h, y, x, x2, ..., x(size-3), Lks2, ...,
# Lks(size-3) and Links2, ..., Links(size-3) (which are only global
# due to recursive procedure calls and are not otherwise used).
global h, Adjhists, Lks2, Links2, histcount, x2, Adjhistsfol, y,
 size, x, Lks3, Links3, x3;
local position, d, Tempset, z;
    if place = 2 then
        member(h, Adjhists[testhist], 'position');
        Lks2 := subsop(position = NULL, Adjhists[testhist]);
        Links2 := sort(convert((convert(Lks2, set) intersect {d $ d =
```

```
        h + 1 .. histcount}), list));
    for x2 from 1 to nops(Links2) do
        if {Links2[x2]} intersect (checkset union
         convert(Adjhists[Adjhistsfol[h][y]], set)) = {} then
            Cycles(3, n, checkset union
             convert(Adjhists[testhist], set), Links2[x2])
        fi
    od
elif place = 3 and size <> 5 then
    member(Adjhistsfol[h][x], Adjhists[testhist], 'position');
    Lks3 := subsop(position = NULL, Adjhists[testhist]);
    Links3 := sort(convert((convert(Lks3, set) intersect {d $ d =
     h + 1 .. histcount}), list));
    for x3 from 1 to nops(Links3) do
        if {Links3[x3]} intersect (checkset union
         convert(Adjhists[Adjhistsfol[h][y]], set)) = {} then
            Cycles(4, n, checkset union
             convert(Adjhists[testhist], set), Links3[x3])
        fi
    od
elif place > 3 and place < (size - 2) then
    member(Links.(place-2)[x.(place-2)], Adjhists[testhist],
     'position');
    Lks.place := subsop(position = NULL, Adjhists[testhist]);
    Links.place := sort(convert((convert(Lks.place, set)
     intersect {d $ d = h + 1 .. histcount}), list));
    for x.place from 1 to nops(Links.place) do
        if {Links.place[x.place]} intersect (checkset union
         convert(Adjhists[Adjhistsfol[h][y]], set)) = {} then
            Cycles(place + 1, n, checkset union
             convert(Adjhists[testhist], set),
             Links.place[x.place])
        fi
```

```
        od
    elif place = (size - 2) then
        Tempset := convert(Adjhists[Adjhistsfol[h][y]], set)
         intersect convert(Adjhists[testhist], set) intersect
         {d $ d = h + 1 .. histcount};
        d := 'd';
        for z from 1 to nops(Tempset) do
            if {Tempset[z]} intersect checkset = {} then
                Cycle.size[cycle.size.no] := [h, Adjhistsfol[h][x],
                 'Links.d[x.d]' $ d = 2 .. size - 3, Tempset[z],
                 Adjhistsfol[h][y]];
                cycle.size.no := cycle.size.no + 1
            fi
        od
    fi
end:

GenerateTypes(r, Kplus1):

qcount := 1:

histno := 1:

Simhists[1] := {}:

histcount := 1:

Storehists[1] := [[], [0]]:

GenerateQs(1, Nplus1, 1):

Qbox := 'Qbox':
```

```
dimtype := (r+K)!/(r!*K!);

histno := histno - 1;

histcount := histcount - 1;

qcount := qcount - 1;

dimq := (Nplus1 + dimtype-1)!/(Nplus1!*(dimtype-1)!);

Links():

AdjacentHists():

GenerateCycles():

readlib(unassign):
unassign('x', 'y', 'h', evaln(Lks.(2 .. size - 3)), evaln(Links.(2 ..
 size - 3)), evaln(x.(2 .. size - 3))):

if size = 3 then
    cycle3no
elif size = 4 then
    cycle4no
else
    cycle.size.no
fi;
```

## A.6 CycleRatios

```
# Assumes that the value of size >= 3 has been assigned and that the
# program FindCycles has been executed.  Reads file probs which is
# assumed to assign the table Probs.  Finds the minimum and maximum
# size-cycle ratio for the probabilities in Probs.

read probs:

EvaluateRatios := proc(size)
# Creates a table called Ratios containing, for each size-cycle, the
# value of its probability ratio for the probabilities in Probs.
global Storehists, Probs, Ratios;
local c, num, den, h, commonq, tail, head, d;
    for c from 1 to cycle.size.no do
        num := 1;
        den := 1;
        for h from 1 to (size - 1) do
            commonq := convert(Storehists[Cycle.size[c][h]][2], set)
             intersect convert(Storehists[Cycle.size[c][h+1]][2],
             set);
            member(op(commonq), Storehists[Cycle.size[c][h]][2],
             'tail');
            member(op(commonq), Storehists[Cycle.size[c][h+1]][2],
             'head');
            num := num*Probs[Cycle.size[c][h], tail];
            den := den*Probs[Cycle.size[c][h+1], head]
        od;
        commonq := convert(Storehists[Cycle.size[c][size]][2], set)
         intersect convert(Storehists[Cycle.size[c][1]][2], set);
        member(op(commonq), Storehists[Cycle.size[c][size]][2],
         'tail');
```

```
        member(op(commonq), Storehists[Cycle.size[c][1]][2], 'head');
        num := num*Probs[Cycle.size[c][size], tail];
        den := den*Probs[Cycle.size[c][1], head];
        Ratios[c] := num/den
    od;
    print(min(Ratios[d] $ d = 1 .. cycle.size.no));
    print(max(Ratios[d] $ d = 1 .. cycle.size.no))
end:

EvaluateRatios(size):
```

# Appendix B

# Data Sets

## B.1 Mosimann's Pollen Data

| *Pinus* | *Abies* | *Quercus* | *Alnus* |
|---|---|---|---|
| 94 | 0 | 5 | 1 |
| 75 | 2 | 14 | 9 |
| 81 | 2 | 13 | 4 |
| 95 | 2 | 3 | 0 |
| 89 | 3 | 1 | 7 |
| 84 | 5 | 7 | 4 |
| 81 | 3 | 10 | 6 |
| 97 | 0 | 2 | 1 |
| 86 | 1 | 8 | 5 |
| 86 | 2 | 11 | 1 |
| 82 | 2 | 10 | 6 |
| 72 | 1 | 16 | 11 |
| 89 | 0 | 9 | 2 |
| 93 | 4 | 2 | 1 |
| 87 | 1 | 11 | 1 |
| 85 | 0 | 12 | 3 |
| continued next page | | | |

| *Pinus* | *Abies* | *Quercus* | *Alnus* |
|---|---|---|---|
| 91 | 0 | 7 | 2 |
| 95 | 1 | 3 | 1 |
| 94 | 3 | 3 | 0 |
| 85 | 1 | 12 | 2 |
| 91 | 1 | 4 | 4 |
| 99 | 1 | 0 | 0 |
| 90 | 2 | 8 | 0 |
| 91 | 0 | 8 | 1 |
| 79 | 1 | 19 | 1 |
| 89 | 0 | 7 | 4 |
| 95 | 2 | 1 | 2 |
| 90 | 3 | 5 | 2 |
| 93 | 1 | 6 | 0 |
| 90 | 2 | 7 | 1 |
| 89 | 2 | 9 | 0 |
| 88 | 1 | 9 | 2 |
| 99 | 0 | 1 | 0 |
| 86 | 1 | 10 | 3 |
| 88 | 0 | 7 | 5 |
| 91 | 0 | 7 | 2 |
| 84 | 0 | 14 | 2 |
| 84 | 1 | 12 | 3 |
| 97 | 0 | 3 | 0 |
| 83 | 0 | 13 | 4 |
| 81 | 1 | 15 | 3 |
| 81 | 1 | 16 | 2 |
| 76 | 2 | 18 | 4 |
| 87 | 3 | 7 | 3 |
| 91 | 1 | 5 | 3 |
| 94 | 0 | 5 | 1 |
| 88 | 1 | 11 | 0 |

continued next page

| *Pinus* | *Abies* | *Quercus* | *Alnus* |
|---|---|---|---|
| 93 | 4 | 2 | 1 |
| 84 | 0 | 8 | 8 |
| 87 | 1 | 12 | 0 |
| 89 | 1 | 6 | 4 |
| 73 | 0 | 13 | 14 |
| 87 | 3 | 8 | 2 |
| 94 | 1 | 3 | 2 |
| 81 | 2 | 9 | 8 |
| 88 | 0 | 9 | 3 |
| 94 | 0 | 4 | 2 |
| 69 | 7 | 18 | 6 |
| 90 | 0 | 8 | 2 |
| 86 | 1 | 8 | 5 |
| 90 | 0 | 7 | 3 |
| 74 | 5 | 16 | 5 |
| 82 | 2 | 11 | 5 |
| 87 | 3 | 9 | 1 |
| 68 | 3 | 26 | 3 |
| 77 | 3 | 11 | 9 |
| 86 | 2 | 7 | 5 |
| 79 | 1 | 11 | 9 |
| 87 | 0 | 11 | 2 |
| 79 | 1 | 17 | 3 |
| 74 | 0 | 19 | 7 |
| 80 | 0 | 14 | 6 |
| 85 | 3 | 9 | 3 |

Table B.1: Forest pollen counts from the Bellas Artes core, Clisby & Sears (1955)