

Multivariate Generalizations of the Multiplicative Binomial Distribution: Introducing the MM Package

P. M. E. Altham
University of Cambridge

Robin K. S. Hankin
University of Stirling

Abstract

We present two natural generalizations of the multinomial and multivariate binomial distributions, which arise from the multiplicative binomial distribution of [Altham \(1978\)](#). The resulting two distributions are discussed and we introduce an R package, **MM**, which includes associated functionality.

This vignette is based on [Altham and Hankin \(2012\)](#).

Keywords: Generalized Linear Models, Multiplicative Binomial, Overdispersion, Overdispersed Binomial, Categorical Exponential Family, Multiplicative Multinomial Distribution, R.

1. Introduction

The uses of the binomial and multinomial distributions in statistical modelling are very well understood, with a huge variety of applications and appropriate software, but there are plenty of real-life examples where these simple models are inadequate. In the current paper, we first remind the reader of the two-parameter exponential family generalization of the binomial distribution first introduced by [Altham \(1978\)](#) to allow for over- or under-dispersion:

$$P(Z = j) = \binom{n}{j} p^j q^{n-j} \theta^{j(n-j)} / f(p, \theta, n) \quad j = 0, \dots, n \quad (1)$$

where

$$f(p, \theta, n) = \sum_{j=0}^n \binom{n}{j} p^j q^{n-j} \theta^{j(n-j)}. \quad (2)$$

Here, $0 \leq p \leq 1$ is a probability, $p + q = 1$, and $\theta > 0$ is the new parameter which controls the shape of the distribution; the standard binomial $\text{Bi}(n, p)$ is recovered if $\theta = 1$. [Altham](#) points out that this distribution is more sharply peaked than the binomial if $\theta > 1$, and more diffuse if $\theta < 1$. As far as we are aware, no other generalization has this type of flexibility; for example, the beta-binomial distribution ([Johnson, Kemp, and Kotz 2005](#)) only allows for *over*-dispersion relative to the corresponding binomial distribution.

We then introduce two different generalizations, both of which are of exponential family form. We call these:

The multivariate multinomial distribution

Take non-negative integers $y_1, \dots, y_k \geq 0$ with $\sum y_i = y$, a fixed integer. Then suppose that the probability mass function of Y_1, \dots, Y_k is

$$P(y_1, \dots, y_k) = C^{-1} \binom{y}{y_1 \dots y_k} \prod_{i=1}^k p_i^{y_i} \prod_{1 \leq i < j \leq k} \theta_{ij}^{y_i y_j} \quad (3)$$

where the free parameters are $\mathbf{p} = (p_1, \dots, p_k)$ and $\boldsymbol{\theta} = \theta_{ij}$, with $p_i \geq 0$ for $1 \leq i \leq k$, $\sum p_i = 1$, and $\theta_{ij} > 0$ for $1 \leq i < j \leq k$ [these restrictions on i, j understood henceforth]. Here $C = C(y, \mathbf{p}, \boldsymbol{\theta})$ is a normalization constant. Thus the standard multinomial is recovered if $\theta_{ij} = 1$, and in this case $C = 1$.

The multivariate multiplicative binomial distribution

For simplicity of notation, we restrict attention to the bivariate case. The proposed frequency function $P(X_1 = x_1, X_2 = x_2) = f(x_1, x_2)$ is

$$f(x_1, x_2) = C^{-1} \binom{m_1}{x_1 z_1} p_1^{x_1} q_1^{z_1} \theta_1^{x_1 z_1} \cdot \binom{m_2}{x_2 z_2} p_2^{x_2} q_2^{z_2} \theta_2^{x_2 z_2} \cdot \phi^{x_1 x_2} \quad (4)$$

where $p_i + q_i = 1$ and $x_i + z_i = m_i$, $i = 1, 2$; all parameters are strictly positive and C is again a normalization constant. Thus X_1, X_2 are independent iff $\phi = 1$. Furthermore, if $\phi = 1$, then $\theta_1 = \theta_2 = 1$ corresponds to X_1, X_2 independent Binomial: $X_i \sim \text{Bi}(m_i, p_i)$.

We then introduce an R ([R Development Core Team 2011](#)) package **MM** which implements some functionality for these distributions. Because of their simple exponential family forms, both these distributions may be fitted to appropriate count data by using the `glm()` R function with the Poisson distribution and log link function. This follows from the ingenious result of [Lindsey and Mersch \(1992\)](#), and has the very important consequence that the computationally expensive normalizing constants in Equations 3 and 4 above need never be evaluated.

Both these distributions are clearly exponential family-type distributions ([Cox and Hinkley 1974](#)), and may be seen as discrete analogues of the multivariate normal distribution in the following two respects, which we illustrate only for the multiplicative multinomial distribution.

Firstly, suppose we have observed frequencies $n(y_1, \dots, y_k)$ with $\sum y_i = y$, then the log likelihood of this dataset may be written as

$$\sum n(y_1, \dots, y_k) \log P(y_1, \dots, y_k). \quad (5)$$

where the summation is over $y_1, \dots, y_k \geq 0$ with $\sum y_i = y$. This gives rather simple expressions, essentially the sample mean and sample covariance matrix of the vector (y_1, \dots, y_k) , for the minimal sufficient statistics of this exponential family. Hence, by standard theory for exponential families, at the maximum likelihood value of the parameters, the observed and fitted values of the minimal sufficient statistics will agree exactly.

Secondly, as we show later, each of the distributions given in Equations 3 and 4 is reproductive under conditioning on components.

1.1. A probabilistic derivation of the Multiplicative Multinomial

It is possible to consider the MM distribution from the perspective of contingency tables, which for simplicity we will carry out for $k = 3$, $y = 4$: The general case is notationally challenging.

Our preferred interpretation is drawn from the field of psephology: Consider a household of 4 indistinguishable voters, each of whom votes for exactly one of 3 political parties, say \wp_1, \wp_2, \wp_3 . Let y_1, y_2, y_3 be the total number of votes obtained from this household for \wp_1, \wp_2, \wp_3 respectively, and so $y_1 + y_2 + y_3 = 4$.

The 4 voters in the household may be considered as corresponding to the rows, columns, layers and the 4th dimension of a $3 \times 3 \times 3 \times 3$ contingency table, with cell probabilities P_{ijhl} for $1 \leq i, j, h, l \leq 3$, which we will assume have the following symmetric form

$$P_{ijhl} = \frac{1}{C'} \cdot p_i p_j p_h p_l \cdot \theta_{ij} \theta_{ih} \theta_{il} \theta_{jh} \theta_{jl} \theta_{hl} \quad (6)$$

where $\theta_{rs} = \theta_{sr}$ for $s, r \in \{i, j, h, l\}$ [notation is analogous to that used in Equation 8 of [Altham \(1978\)](#), with θ written for ϕ], and without loss of generality $\theta_{rr} = 1$.

The parameters θ_{ij} may be interpreted in terms of *conditional* cross-ratios. Recalling that $P_{ijhl} = P_{ijlh} = \dots = P_{lhji}$ we have, for example:

$$\frac{P_{12hl} P_{21hl}}{P_{11hl} P_{22hl}} = \theta_{12}^2 \quad \text{for each } h, l. \quad (7)$$

By enumerating the possible voting results for a given family of size 4, we may find the resulting joint distribution of (Y_1, Y_2, Y_3) , where random variable Y_i is the household total of votes for party \wp_i , $i = 1, 2, 3$. For example, $P(Y_1 = 4, Y_2 = 0, Y_3 = 0) = \frac{1}{C'} \cdot p_1^4$ is the probability that all 4 members of the household vote for \wp_1 . Similarly, $P(Y_1 = 3, Y_2 = 1, Y_3 = 0) = \frac{1}{C'} \cdot 4p_1^3 p_2^1 \theta_{12}^3$ is the probability that 3 members of the household vote for \wp_1 and the remaining 1 member votes for \wp_2 .

This clearly corresponds to the given Multiplicative Multinomial distribution, so $C = C'$. We return to this example with a synthetic dataset in Section 3.1 below.

1.2. Marginal and conditional distributions

There does not appear to be an elegant expression for the marginal distribution of (Y_1, \dots, Y_r) where $r < k$. However, the multiplicative multinomial behaves ‘elegantly’ under conditioning on a subset of the variables (Y_1, \dots, Y_k) . For example,

$$P(y_1, y_2 | y_3, \dots, y_k) \propto \frac{\phi_1^{y_1} \phi_2^{y_2} \theta_{12}^{y_1 y_2}}{y_1! y_2!}, \quad y_1 + y_2 = y - \sum_{i=3}^k y_i \quad (8)$$

where

$$\phi_1 = p_1 \theta_{13}^{y_3} \dots \theta_{1k}^{y_k} \quad \phi_2 = p_2 \theta_{23}^{y_3} \dots \theta_{2k}^{y_k}. \quad (9)$$

Hence the distribution of Y_1 , conditional on the values of (y_3, \dots, y_k) is multiplicative binomial in the sense of [Altham \(1978\)](#). Similarly, the distribution of (Y_1, \dots, Y_ν) , conditional on the values of $(y_{\nu+1}, \dots, y_k)$ is multiplicative multinomial in the sense defined above.

1.3. The normalization constant

The constant C in Equation 3 must be determined by numerical means:

$$C = \sum_{y_1 + \dots + y_k = y} \binom{y}{y_1 \dots y_k} \prod_{i=1}^k p_i^{y_i} \prod_{1 \leq i < j \leq k} \theta_{ij}^{y_i y_j}. \quad (10)$$

Although this is provided in the **MM** package [function `NormC()`], it is computationally expensive and difficult to evaluate for all but small k and y .

1.4. The Poisson method for fitting distributions with an intractable normalizing constant

The parameters (\mathbf{p}, θ) may be estimated without determining the normalizing constant C by transforming the problem into a Generalized Linear Model. The method presented here follows [Lindsey and Mersch \(1992\)](#); for simplicity of notation we take $k = 3$. Equation 3 is equivalent to

$$\log P(y_1, y_2, y_3) = \mu + \sum y_i \log p_i + \sum_{i < j} y_i \cdot y_j \log \theta_{ij} + \mathbf{offset}[y_1, y_2, y_3] \quad (11)$$

where $\mathbf{offset}[y_1, y_2, y_3] = -\sum \log(y_i!)$ accounts for the multinomial term on the right hand side. The log-likelihood \mathcal{L} of the dataset $n(y_1, y_2, y_3)$ is given by

$$\mathcal{L} = \sum_{y_1 + y_2 + y_3 = y} n(y_1, y_2, y_3) \log P(y_1, y_2, y_3). \quad (12)$$

Thus, treating $n(y_1, y_2, y_3)$ as independent Poisson variables with parameters¹ given by Equation 11, we may fit the parameters of Equation 3 using `glm(..., family=poisson)`, using the canonical log link function, and regressing $n(y_1, y_2, y_3)$ on the variables

$$y_1, \quad y_2, \quad y_3, \quad y_1 y_2, \quad y_1 y_3, \quad y_2 y_3.$$

With obvious notation, the R idiom is

```
glm(n ~ -1+offset(Off) + y1      + y2      + y3
      + y1:y2 + y1:y3 + y2:y3,
     family = poisson)
```

(recall that $y_1 + y_2 + y_3 = y$, fixed), which is given by function `Lindsey()` in the package.

¹The distribution of independent Poisson random variables conditional on their total is multinomial with probabilities equal to the scaled Poisson parameters. If $X_i \sim \text{Po}(\lambda_i)$, then elementary considerations show

$$P\left(X_1 = x_1, \dots, X_k = x_k \mid \sum_i x_i = N\right) = \binom{N}{x_1, \dots, x_k} \prod \left(\frac{\lambda_i}{\sum_i \lambda_i}\right)^{x_i},$$

the right hand side being recognisable as a multinomial distribution. Given that the distribution is of the exponential family, it is the case that $\sum n = \sum \hat{\lambda}_i$, the normalizing constant is not needed.

1.5. Multivariate multiplicative binomial

Considering the bivariate case for simplicity, suppose (X_1, X_2) to be non-negative integers not exceeding known fixed maxima m_1, m_2 respectively.

We introduce a 5-parameter distribution of exponential family form. In common with the multiplicative binomial, it has the property that at the maximum likelihood values of these parameters, the observed and fitted values of the means of X_i , $i = 1, 2$ will agree exactly, and similarly for the observed and fitted values of the covariance matrix of X_1, X_2 . This distribution is easy to fit to frequency data (again using the [Lindsey and Mersch](#) Poisson device). The distribution has some nice properties, but there do not appear to be simple formulæ for its moments.

The proposed frequency function $P(X_1 = x_1, X_2 = x_2) = f(x_1, x_2)$ is

$$f(x_1, x_2) = C^{-1} \binom{m_1}{x_1 \ z_1} p_1^{x_1} q_1^{z_1} \theta_1^{x_1 z_1} \cdot \binom{m_2}{x_2 \ z_2} p_2^{x_2} q_2^{z_2} \theta_2^{x_2 z_2} \cdot \phi^{x_1 x_2} \quad (13)$$

where $p_i + q_i = 1$ and $x_i + z_i = m_i$; all parameters are strictly positive.

Here, C is the normalization constant:

$$C = \sum_{x_1+z_1=m_1} \sum_{x_2+z_2=m_2} \binom{m_1}{x_1 \ z_1} p_1^{x_1} q_1^{z_1} \theta_1^{x_1 z_1} \cdot \binom{m_2}{x_2 \ z_2} p_2^{x_2} q_2^{z_2} \theta_2^{x_2 z_2} \cdot \phi^{x_1 x_2} \quad (14)$$

Thus X_1, X_2 are independent iff $\phi = 1$. As already noted, if $\phi = 1$, then $\theta_1 = \theta_2 = 1$ corresponds to X_1, X_2 independent Binomial: $X_i \sim \text{Bi}(m_i, p_i)$.

Although there does not seem to be a simple expression for the correlation between X_1 and X_2 , it is easily seen that ϕ controls their interdependence in a likelihood ratio fashion, with

$$\frac{f(x_1, x_2) f(x_1 + 1, x_2 + 1)}{f(x_1 + 1, x_2) f(x_1, x_2 + 1)} = \phi. \quad (15)$$

Indeed, following [Lehmann \(1966\)](#) we can prove a much stronger statement: The variables X_1, X_2 are positive monotone likelihood ratio dependent² for $\phi > 1$, negative if $\phi < 1$. The conditional distribution is again of multiplicative binomial form, since we can write

$$P(X_1 = x_1 | X_2 = x_2) \propto \binom{m_1}{x_1 \ z_1} (p_1 \phi^{x_2})^{x_1} q_1^{z_1} \theta_1^{x_1 z_1}. \quad (16)$$

2. The MM package

The **MM** package associated with this article provides R functionality for assessing the multiplicative multinomial and multivariate binomial. We have provided user-friendly wrappers to expedite use of the distributions in a data analysis setting.

²Random variables X_1, X_2 are positive monotone likelihood ratio dependent if $f(x_1, x_2') f(x_1', x_2) \leq f(x_1, x_2) f(x_1', x_2')$ for all $x_1 < x_1', x_2 < x_2'$, and negative monotone likelihood ratio dependent if the inequality is reversed.

The **MM** package uses an object-oriented approach: The set of free parameters (one vector, one upper-diagonal matrix) is not a standard R object, and is defined to be an object of S4 class **paras**. The objects thus defined are user-transparent and a number of manipulation methods are provided in the package.

For example, consider Equation 3 with $k = 5$ and $p_i = \frac{1}{5}$, $1 \leq i \leq 5$ and $\theta_{ij} = 2$ for $1 \leq i < j \leq 5$. This distribution would be underdispersed compared with the corresponding multinomial. It is straightforward to create an object corresponding to the parameters for this distribution using the package:

```
R> library("MM")
R> pm1 <- paras(5, pnames=letters[1:5])
R> theta(pm1) <- 2
R> pm1
```

```
$p
  a    b    c    d    e
0.2 0.2 0.2 0.2 0.2
```

```
$theta
  a  b  c  d  e
a NA  2  2  2  2
b NA NA  2  2  2
c NA NA NA  2  2
d NA NA NA NA  2
e NA NA NA NA NA
```

Now we may sample repeatedly from the distribution (sampling is quick because it does not require evaluation of the normalization constant). Consider $y = 20$:

```
R> set.seed(0)
R> (sample1 <- rMM(n=10, Y=20, paras=pm1))
```

```
      a b c d e
[1,] 4 4 4 5 3
[2,] 5 3 4 5 3
[3,] 4 5 4 4 3
[4,] 4 4 4 4 4
[5,] 4 3 6 3 4
[6,] 4 4 3 4 5
[7,] 4 4 5 3 4
[8,] 5 4 4 3 4
[9,] 3 6 3 4 4
[10,] 4 4 5 4 3
```

See how closely clustered the sample is around its mean of (4, 4, 4, 4, 4); compare the wider dispersion of the multinomial:

```
R> pm2 <- pm1
R> theta(pm2) <- 1 # revert to classical multinomial
R> (sample2 <- rMM(n=10, Y=20, paras=pm2))
```

```
      a b c d e
[1,] 5 5 1 5 4
[2,] 6 6 1 4 3
[3,] 4 1 4 8 3
[4,] 5 3 6 4 2
[5,] 3 3 2 4 8
[6,] 2 4 3 4 7
[7,] 6 2 3 3 6
[8,] 3 3 6 4 4
[9,] 1 6 2 8 3
[10,] 5 3 2 5 5
```

Thus `sample2` is drawn from the classical multinomial. It is then straightforward to perform a likelihood ratio test on, say, `sample1`:

```
R> support1 <- MM_allsamesum(sample1, paras=pm1)
R> support2 <- MM_allsamesum(sample1, paras=pm2)
```

```
R> support1-support2
```

```
[1] 17.335
```

Function `MM_allsamesum()` calculates the log likelihood for a specific parameter object (in this case, `pm1` and `pm2` respectively) and we see that, for `sample1`, hypothesis `pm1` is preferable to `pm2` on the grounds of a likelihood ratio of about $\Lambda = 0.03 \times 10^{-6}$, corresponding to 17.33 units of support. This would exceed the two units of support criterion suggested by [Edwards \(1992\)](#) and we could reject `pm2`. Alternatively, we could observe that $-2 \log \Lambda$ is in the tail region of its asymptotic distribution, χ^2_1 .

The package includes a comprehensive suite of functionality which is documented through the R help system and accessible by typing `library(help="MM")` at the command prompt.

3. The package in use

The package comes with a number of datasets, four of which are illustrated here. We begin with a small synthetic dataset, then consider data taken from the social sciences, previously analyzed by [Wilson \(1989\)](#); analyze some pollen counts considered by [Mosimann \(1962\)](#) in the context of palaeoclimatology; and finally assess a marketing science dataset.

3.1. Synthetic voting dataset

We begin with a small synthetic dataset which is simple enough to illustrate the salient aspects of the multiplicative multinomial distribution, and the **MM** package.

This dataset arises from 96 households each of size 4, in which each member of the household is noted as voting Lib, Con or Lab respectively. We take $n(\cdot, \cdot, \cdot)$ as the voting tally; thus $n(0, 0, 4) = 5$ [the first line] means that there are exactly 5 households in which all 4 members vote Labour; similarly $n(0, 1, 3) = 8$ means that there are exactly 8 households in which 1 member votes Conservative and the remaining 3 vote Labour.

```
R> data("voting")
R> cbind(voting, voting_tally)
```

	Lib	Con	Lab	voting_tally
[1,]	0	0	4	5
[2,]	0	1	3	8
[3,]	0	2	2	7
[4,]	0	3	1	4
[5,]	0	4	0	6
[6,]	1	0	3	1
[7,]	1	1	2	7
[8,]	1	2	1	4
[9,]	1	3	0	9
[10,]	2	0	2	5
[11,]	2	1	1	7
[12,]	2	2	0	12
[13,]	3	0	1	2
[14,]	3	1	0	7
[15,]	4	0	0	12

One natural hypothesis is that the data are drawn from a multinomial distribution (alternative hypotheses might recognize that individuals within a given household may be non-independent of each other in their voting).

The multinomial hypothesis may be assessed using `glm()` following [Lindsey and Mersch \(1992\)](#) but without the interaction terms:

```
R> Off <- -rowSums(lfactorial(voting))
R> summary(glm(voting_tally ~ -1 + (.) + offset(Off),
               data = as.data.frame(voting), family = poisson))
```

Call:

```
glm(formula = voting_tally ~ -1 + (.) + offset(Off), family = poisson,
     data = as.data.frame(voting))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
Lib	0.9548	0.0706	13.51	< 2e-16
Con	0.9122	0.0728	12.54	< 2e-16
Lab	0.6099	0.0886	6.88	5.8e-12

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 543.236 on 15 degrees of freedom
 Residual deviance: 77.315 on 12 degrees of freedom
 AIC: 137

Number of Fisher Scoring iterations: 5

Thus the model fails to fit (77.315 being much larger than the corresponding degrees of freedom, 12). This is because the observed frequencies of the cells in which all members of the household vote for the same party (namely for rows 1, 5 and 15 of the data) greatly exceed the corresponding expected numbers under the simple multinomial model.

The next step is to take account of the fact that individuals within a given household may be non-independent of each other in their voting intentions (and may indeed tend to disagree with each other rather than all vote the same way). Positive dependence between individuals in a household could be modelled by the Dirichlet-multinomial distribution ([Mosimann 1962](#)), but by using the Multiplicative Multinomial introduced here, we are allowing dependence between individuals in a household to be positive or negative.

The MM parameters may be estimated, again following [Lindsey and Mersch \(1992\)](#) but this time admitting first-order interaction, using bespoke function `Lindsey()`:

```
R> Lindsey(voting, voting_tally, give_fit = TRUE)
```

\$MLE

\$p

	Lib	Con	Lab
	0.36695	0.31515	0.31790

\$theta

		Lib	Con	Lab
Lib	NA	0.67351	0.48259	
Con	NA		NA	0.65153
Lab	NA		NA	NA

\$fit

Call:

```
glm(formula = jj$d ~ -1 + offset(Off) + (.)^2, family = poisson,
     data = data.frame(jj$tbl))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
Lib	1.3745	0.0712	19.31	< 2e-16
Con	1.2224	0.0898	13.61	< 2e-16
Lab	1.2311	0.0925	13.31	< 2e-16

```

Lib:Con  -0.3952      0.0844   -4.68   2.8e-06
Lib:Lab  -0.7286      0.1023   -7.13   1.0e-12
Con:Lab  -0.4284      0.0965   -4.44   8.9e-06

```

(Dispersion parameter for poisson family taken to be 1)

```

Null deviance: 543.236  on 15  degrees of freedom
Residual deviance: 11.501  on 9  degrees of freedom
AIC: 77.16

```

Number of Fisher Scoring iterations: 5

Observe that the MLEs of p [viz (0.367, 0.315, 0.318)] are obtained as proportional to the exponential of the estimated regression coefficients: $(e^{1.375}, e^{1.222}, e^{1.231})$, normalized to add to 1.

This model is quite a good fit in the sense that the null deviance 11.5 is not in the tail region of χ_9^2 , its null distribution; it can be seen that all 3 interaction parameters are significant and, for example, $\hat{\theta}_{23} = 0.652 = \exp(-0.428)$.

The corresponding conditional cross-ratios are all significantly greater than 1; for example

$$\frac{\hat{P}_{11hl}\hat{P}_{22hl}}{\hat{P}_{12hl}\hat{P}_{21hl}} = \frac{1}{\hat{\theta}_{12}^2} = \frac{1}{0.6735116^2} = 2.2045. \quad (17)$$

3.2. Housing satisfaction data

We now consider a small dataset taken from Table 1 of [Wilson \(1989\)](#), who analyzed the dataset in the context of overdispersion ([Wilson](#) himself took the dataset from Table 1 of [Brier \(1980\)](#)).

In a non-metropolitan area, there were 18 independent neighbourhoods each of 5 households, and each household gave its response concerning its personal satisfaction with their home. The allowable responses were ‘unsatisfied’ (US), ‘satisfied’ (S), and ‘very satisfied’ (VS).

```

R> data("wilson")
R> head(non_met)

```

```

      US S VS
[1,]  3 2  0
[2,]  3 2  0
[3,]  0 5  0
[4,]  3 2  0
[5,]  0 5  0
[6,]  4 1  0

```

Thus the first neighbourhood had three households responding US, two reporting S, and zero reporting VS; the second neighbourhood had the same reporting pattern.

Observe that the 5 households within a neighbourhood may not be independent in their responses. The first step is to recast the dataset into a table format; the package provides a function `gunter()` [named for the R-lister who suggested the elegant and fast computational method]:

```
R> wilson <- gunter(non_met)
R> wilson
```

```
$tbl
      US S VS
1      5 0 0
2      4 1 0
3      3 2 0
4      2 3 0
5      1 4 0
6      0 5 0
7      4 0 1
8      3 1 1
9      2 2 1
10     1 3 1
11     0 4 1
12     3 0 2
13     2 1 2
14     1 2 2
15     0 3 2
16     2 0 3
17     1 1 3
18     0 2 3
19     1 0 4
20     0 1 4
21     0 0 5

$d
[1] 1 5 4 2 0 2 1 0 0 1 1 0 0 1 0 0 0 0 0 0 0
```

Thus 1 neighbourhood reported $c(5,0,0)$, and 5 neighbourhoods reported $c(4,1,0)$ [because $d[1]=1$ and $tbl[1,]=c(5,0,0)$; and $d[2]=5$ and $tbl[2,]=c(4,1,0)$ respectively].

The hypothesis that the data are drawn from a multinomial distribution may again be assessed by using [Lindsey and Mersch's](#) technique:

```
R> attach(wilson)
R> Off <- -rowSums(lfactorial(tbl))
R> summary(glm(d~-1+(.)+offset(Off),data=tbl,family=poisson))
```

Call:

```
glm(formula = d ~ -1 + (.) + offset(Off), family = poisson, data = tbl)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
US	0.886	0.111	7.96	1.7e-15
S	0.673	0.132	5.10	3.4e-07
VS	-1.355	0.437	-3.10	0.0019

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 110.17 on 21 degrees of freedom
 Residual deviance: 21.22 on 18 degrees of freedom
 AIC: 49.19

Number of Fisher Scoring iterations: 6

Thus the multinomial model is a reasonable fit, in the sense that the residual deviance of 21.22 is consistent with the null distribution, χ^2_{18} . The slightly increased value would be because the observed frequencies for neighbourhoods in agreement (that is, either perfect agreement— $c(5,0,0)$ or $c(0,5,0)$ or $c(0,0,5)$ —or near-perfect, as in $c(4,1,0)$) exceed the corresponding expected numbers under the simple multinomial model.

The MM parameters may be estimated, again following [Lindsey and Mersch \(1992\)](#) but this time admitting first-order interaction:

```
R> Lindsey(wilson,give_fit=TRUE)
```

\$MLE

\$p

	US	S	VS
	0.494469	0.411254	0.094277

\$theta

	US	S	VS
US	NA	0.74424	0.59647
S	NA	NA	0.88449
VS	NA	NA	NA

\$fit

Call:

```
glm(formula = jj$d ~ -1 + offset(Off) + (.)^2, family = poisson,
     data = data.frame(jj$tbl))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
US	1.140	0.112	10.20	< 2e-16

S	0.956	0.161	5.95	2.6e-09
VS	-0.517	1.652	-0.31	0.7543
US:S	-0.295	0.112	-2.64	0.0084
US:VS	-0.517	0.508	-1.02	0.3086
S:VS	-0.123	0.545	-0.23	0.8219

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 110.173 on 21 degrees of freedom
 Residual deviance: 13.425 on 15 degrees of freedom
 AIC: 47.4

Number of Fisher Scoring iterations: 6

Thus in this dataset, only the first interaction parameter **US:S** is significant. This might be interpreted as an absence of **VS** responses coupled with a broader than expected spread split between **US** and **S**. Note that the residual deviance is now less than the corresponding degrees of freedom.

In this case, the three categories **US**, **S**, and **VS** are ordered, a feature which is not used in the present approach. It is not clear at this stage how we could best include information about such ordering into our analysis.

We now check agreement of the observed and expected sufficient statistics:

```
R> summary(suffstats(wilson))
```

```
$row_sums
      US      S      VS
2.61111 2.11111 0.27778

$cross_prods
      US      S      VS
US 9.27778 3.38889 0.38889
S  3.38889 6.55556 0.61111
VS 0.38889 0.61111 0.38889
```

The `summary()` method gives normalized statistics so that, for example, the `row_sums` total $y = 5$. This may be compared with the expectation of the maximum likelihood MM distribution:

```
R> L <- Lindsey(wilson)

R> expected_suffstats(L,5)

$row_sums
      US      S      VS
2.61111 2.11111 0.27778
```

```
$cross_prods
      US      S      VS
US 9.27778 3.38889 0.38889
S  3.38889 6.55556 0.61111
VS 0.38889 0.61111 0.38889
```

showing agreement to within numerical precision.

3.3. Mosimann's forest pollen dataset

Palynology offers a unique perspective on palaeoclimate; pollen is durable, easily identified, and informative about climate (Faegri and Iversen 1992). We now consider a dataset collected in the context of palaeoclimate (Sears and Clisby 1955; Clisby and Sears 1955), and further analyzed by Mosimann (1962).

We consider a dataset taken from the Bellas Artes core from the Valley of Mexico (Clisby and Sears 1955, Table 2); details of the site are given by Foreman (1955). The dataset comprises a matrix with $N = 73$ observations, each representing a depth in the core, and $k = 4$ columns, each representing a different type of pollen. We follow Mosimann in assuming that the 73 observations are independent, and in restricting the analysis to depths at which a full complement of 100 grains were identified.

```
R> data("pollen")
R> pollen <- as.data.frame(pollen)
R> head(pollen)
```

	Pinus	Abies	Quercus	Alnus
1	94	0	5	1
2	75	2	14	9
3	81	2	13	4
4	95	2	3	0
5	89	3	1	7
6	84	5	7	4

Thus each row is constructed to sum to 100, and there are 4 distinct types of pollen; hence in our notation $y = 100$ and $k = 4$.

Observe that this dataset, in common with the housing satisfaction data considered above, has to be coerced to histogram form; but this time the numbers are larger. The **partitions** package (Hankin 2006) uses generating functions to determine that there are exactly

```
R> library("partitions")
```

```
[1] "partitions" "MM"          "stats"       "graphics"    "grDevices"
[6] "utils"      "datasets"    "methods"     "base"
```

```
R> S(rep(100,4),100)
```

[1] 176851

possible non-negative integer solutions to $y_1 + y_2 + y_3 + y_4 = 100$ (most of these have zero observed count). Each of these solutions must be generated and this is achieved using the `compositions()` function of the **partitions** package ([Hankin 2007](#)).

First we repeat some of [Mosimann](#)'s calculations as a check. Using the ordinary multinomial distribution $Mn(y, p)$, we find \hat{p} to be

```
R> p.hat <- colSums(pollen)/sum(pollen)
```

Pinus	Abies	Quercus	Alnus
0.8627	0.0141	0.0907	0.0325

The observed sample variances for the counts are

```
R> apply(pollen, 2, var)
```

Pinus	Abies	Quercus	Alnus
48.51	2.08	25.87	8.19

but if the ordinary multinomial model held, we would expect these variances to be

```
R> 100*p.hat*(1-p.hat)
```

Pinus	Abies	Quercus	Alnus
11.84	1.39	8.25	3.14

respectively. This shows that the dataset has pronounced over-dispersion compared to the ordinary multinomial. Furthermore, the sample correlation matrix is not what we would expect from the ordinary multinomial.

As [Mosimann](#) points out, the sample correlation matrix is

```
R> cor(pollen)
```

	Pinus	Abies	Quercus	Alnus
Pinus	1.000	-0.3018	-0.896	-0.6892
Abies	-0.302	1.0000	0.087	0.0761
Quercus	-0.896	0.0870	1.000	0.3595
Alnus	-0.689	0.0761	0.360	1.0000

while the correlation matrix for the multinomial corresponding to \hat{p} is actually

	Pinus	Abies	Quercus	Alnus
Pinus	1.000	-0.2999	-0.7917	-0.4592
Abies	-0.300	1.0000	-0.0378	-0.0219
Quercus	-0.792	-0.0378	1.0000	-0.0578
Alnus	-0.459	-0.0219	-0.0578	1.0000

(We have corrected what seems to be a small typo in the 4th column of this matrix in Mosimann's Table 2). It is particularly striking that the data show *positive* correlations for 3 entries. Such positive correlations could never arise from the Dirichlet-multinomial distribution, but they will be exactly matched by our new multiplicative multinomial model. The full sample covariance matrix for the dataset is

```
R> var(pollen)
```

	Pinus	Abies	Quercus	Alnus
Pinus	48.51	-3.031	-31.741	-13.735
Abies	-3.03	2.079	0.638	0.314
Quercus	-31.74	0.638	25.870	5.233
Alnus	-13.74	0.314	5.233	8.188

which is precisely the covariance of the multiplicative multinomial distribution at the maximum likelihood (ML) parameters.

Calculating the Normalizing constant for the MM distribution is computationally expensive; `NormC()` takes over 60 seconds to execute on a 2.66 GHz Intel PC running linux. For direct maximization of the log-likelihood function, for example by **MM** function `optimizer()`, one would have to call `NormC()` many times. Thus function `Lindsey()` represents, in this case, a considerable saving of time in maximizing the log-likelihood (the call below took under 15 seconds elapsed time):

```
Lindsey(pollen, give_fit=TRUE)
```

```
$MLE
```

```
$p
```

Pinus	Abies	Quercus	Alnus
0.273	0.264	0.222	0.241

```
$theta
```

	Pinus	Abies	Quercus	Alnus
Pinus	NA	0.955	0.973	0.954
Abies	NA	NA	0.962	0.954
Quercus	NA	NA	NA	0.975
Alnus	NA	NA	NA	NA

```
$fit
```

```
Call:
```

```
glm(formula = jj$d ~ -1 + offset(Off) + (. )^2, family = poisson,
     data = data.frame(jj$tbl))
```

```
Coefficients:
```

```
Estimate Std. Error z value Pr(>|z|)
```

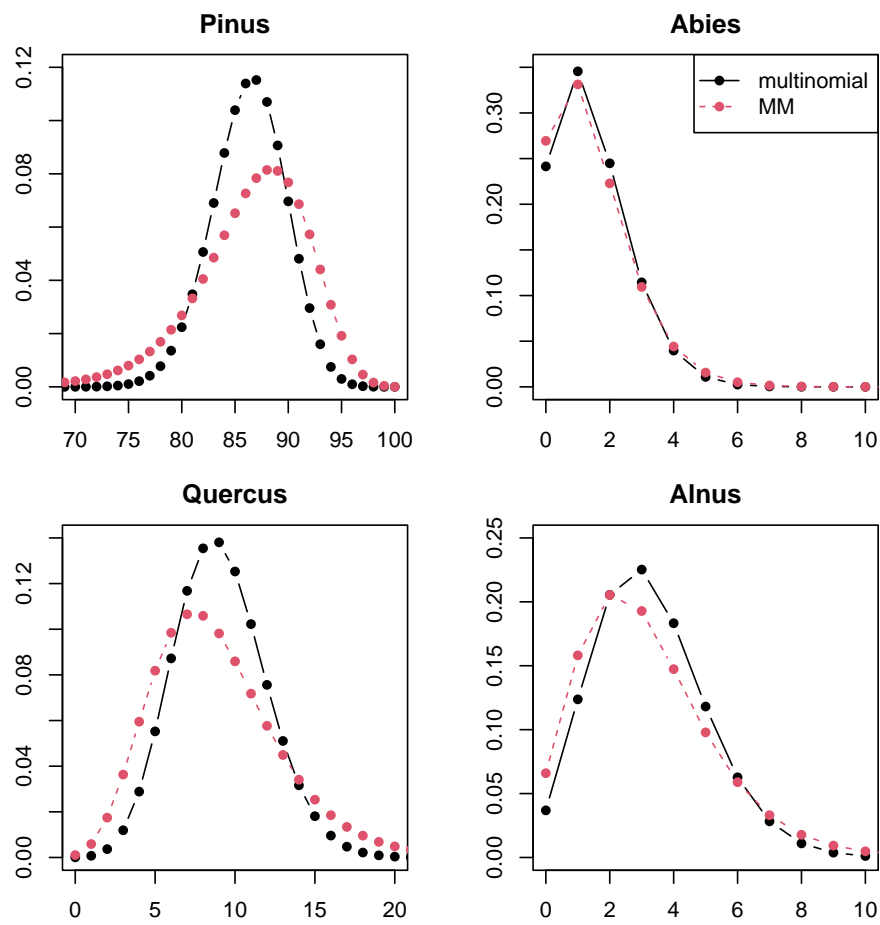



Figure 1: Marginal frequency distributions for numbers of each of four pollen types based on the multinomial distribution (black) and the multiplicative multinomial (red).

Pinus	3.58095	0.00434	825.28	< 2e-16
Abies	3.55110	0.18558	19.13	< 2e-16
Quercus	3.37725	0.11641	29.01	< 2e-16
Alnus	3.45699	0.13275	26.04	< 2e-16
Pinus:Abies	-0.04638	0.00260	-17.81	< 2e-16
Pinus:Quercus	-0.02784	0.00133	-20.87	< 2e-16
Pinus:Alnus	-0.03918	0.00134	-29.23	< 2e-16
Abies:Quercus	-0.04721	0.01661	-2.84	0.0045
Abies:Alnus	-0.04749	0.03123	-1.52	0.1284
Quercus:Alnus	-0.02563	0.00585	-4.38	1.2e-05

(Dispersion parameter for poisson family taken to be 1)

```

Null deviance: 5127.46 on 176851 degrees of freedom
Residual deviance: 402.95 on 176841 degrees of freedom
AIC: 563.4

```

```
Number of Fisher Scoring iterations: 23
```

Thus we arrive at an apparently rather symmetrical set of parameter estimates (in the sense of the elements of \hat{p} being close to one another, and the elements of $\hat{\theta}$ being close to unity). For this dataset, we observe that the asymptotic distribution of the residual deviance, χ^2_{176841} , is not a good approximation for its actual distribution. This is because the frequency data is overwhelmingly comprised of zeros, with only 66 nonzero frequencies amongst the 176851 compositions. Function `glm()` tenders a warning to this effect.

3.4. Marketing science: An example of the multivariate multiplicative binomial

We now illustrate the multivariate multiplicative binomial with an example drawn from the field of economics. [Danaher and Hardie \(2005\)](#) considered a dataset obtained from a sample of $N = 548$ households over four consecutive store trips. For each household, they counted the total number of egg purchases in their four eligible shopping trips, and the total number of bacon purchases for the same trips; the hypothesis was that egg consumption was correlated with bacon consumption.

The dataset is provided with the **MM** package:

```

R> data("danaher")
R> danaher

```

```

      eggs
bacon  0   1   2   3   4
0 254 115  42  13   6
1   34   29  16   6   1
2    8    8   3   3   1
3    0    0   4   1   1
4    1    1   1   0   0

```

Thus 16 households purchased eggs twice and bacon once (this is [Danaher and Hardie's](#) Table 1). The purchases of eggs and bacon are not independent³ and we suggest fitting this data to the distribution given in Equation 4; here $m_1 = m_2 = 4$. The Poisson device of [Lindsey and Mersch](#) is again applicable:

```
R> fit <- Lindsey_MB(danaher)
R> summary(fit)
```

Call:

```
glm(formula = d ~ (.), family = poisson, data = x, offset = Off)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.5117	0.0606	90.92	<2e-16
bacon	-1.6528	0.1667	-9.92	<2e-16
eggs	-1.0487	0.0810	-12.94	<2e-16
`bacon:nbacon`	-0.5149	0.0627	-8.22	<2e-16
`eggs:neggs`	-0.3555	0.0406	-8.76	<2e-16
`bacon:eggs`	0.3006	0.0598	5.03	5e-07

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 3046.440 on 24 degrees of freedom
Residual deviance: 18.666 on 19 degrees of freedom
AIC: 108.3
```

Number of Fisher Scoring iterations: 5

and `glm()` gives a good fit in the sense that the residual deviance of 18.666 is compatible with its asymptotic null distribution χ^2_{19} .

The `bacon:eggs` coefficient (ie $\log \hat{\phi} = 0.3006$) gives $\hat{\phi} = e^{0.3006} = 1.3507$, showing strong positive association.

We can now verify that the expected (marginal) number of egg purchases and bacon purchases under the ML distribution match the observed. The first step is to create `M`, the expected contingency matrix:

```
R> M <- danaher
R> M[] <- fitted.values(fit)
R> M
```

	eggs				
bacon	0	1	2	3	4
0	247.566	119.448	43.999	14.665	3.732
1	40.470	26.374	13.122	5.907	2.030

³Fisher's exact test gives a p value of 11×10^{-6} with 10^6 replicates.

2	6.947	6.115	4.109	2.499	1.160
3	1.484	1.765	1.602	1.315	0.825
4	0.333	0.535	0.656	0.727	0.616

Then we may verify, for example, that the fitted sum of bacon purchases matches its observed value:

```
R> bacon <- slice.index(danaher,1)
R> eggs  <- slice.index(danaher,2)

R> sum(bacon*danaher)          # Observed number of bacon purchases

[1] 710

R> sum(bacon*M)                # Expectation; matches

[1] 710
```

As a final check, we can verify that the sample covariance matches the distribution's covariance at the MLE:

```
R> sum(bacon*eggs*danaher)/N - sum(bacon*danaher)*sum(eggs*danaher)/N^2

[1] 0.144

R> sum(bacon*eggs*M)          /N - sum(bacon*M) *sum(eggs*M) /N^2

[1] 0.144
```

again showing agreement to within numerical precision.

4. Suggestions for further work

The multiplicative multinomial is readily generalized to a distribution with $2^k - 1$ parameters:

$$P(y_1, \dots, y_k) = \binom{y}{y_1 \dots y_k} \prod_{S \subseteq [k]} (\Theta_S)^{\prod_{i \in S} y_i} \quad (18)$$

where $[k] = \{1, 2, \dots, k\}$ is the set of all strictly positive integers not exceeding k . Here, the parameters are indexed by a subset of $[k]$; it is interesting to note that Θ_\emptyset formally represents the normalization constant C . In this notation, Equation 3 becomes

$$\prod_{\substack{S \subseteq [k] \\ |S| \leq 2}} \binom{y}{y_1 \dots y_k} (\Theta_S)^{\prod_{i \in S} y_i} . \quad (19)$$

Equation 18 leads to a distribution of the exponential family type; but interpretation of the parameters is difficult, and further work would be needed to establish the usefulness of this extension.

Further, Equation 4 generalizes to

$$\prod_{i=1}^t \binom{m_i}{x_i z_i} p_i^{x_i} q_i^{z_i} \theta_i^{x_i z_i} \prod_{i < j} \phi_{ij}^{x_i x_j}. \quad (20)$$

It is possible to generalize the equations in a slightly different way. Consider an $r \times c$ matrix \mathbf{n} with entries n_{ij} and fixed marginal totals. Now suppose that each row of \mathbf{n} comprises independent observations from a multinomial distribution with probabilities p_1, \dots, p_r , and likewise the columns are multinomial q_1, \dots, q_c : This is the null of Fisher's exact test. Then one natural probability measure would be

$$P(\mathbf{n}) = \frac{1}{C} \cdot \frac{\prod_{1 \leq i_1 < i_2 \leq r} \theta_{i_1 i_2}^{\sum_{j=1}^c n_{i_1 j} n_{i_2 j}} \prod_{1 \leq j_1 < j_2 \leq c} \phi_{j_1 j_2}^{\sum_{i=1}^r n_{i j_1} n_{i j_2}}}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!} \quad (21)$$

(the fixed known row- and column- sums mean that p_i and q_j , and the marginal multinomial terms, are absorbed into the normalizing constant C). With a slight abuse of notation this can be written

$$P(\mathbf{n}) = \frac{1}{C} \cdot \frac{\prod \Theta^{\mathbf{n}\mathbf{n}^\top} \prod \Phi^{\mathbf{n}^\top \mathbf{n}}}{\prod \mathbf{n}!} \quad (22)$$

where Θ governs row-wise departures from multinomial and Φ governs column-wise departures; there are a total of $r(r-1)/2 + c(c-1)/2$ free parameters.

The Poisson device of [Lindsey and Mersch](#) is again applicable, with the difference that the enumeration carried out by `compositions()` is replaced by enumeration of contingency tables with the correct marginal totals: Function `allboards()` of the `aylmer` package ([West and Hankin 2008](#)). A simple example is given under `help(sweets)`.

As pointed out by an anonymous referee, it might be possible to extend either or both of the new distributions to the context of regression on covariates.

5. Conclusions

In this paper, we considered natural generalizations of the multiplicative binomial distribution to the multivariate case. The resulting distributions have a number of desirable features, including a more precise control over the variance than the multinomial, and a straightforward interpretation in terms of contingency tables.

The distributions belong to the exponential family; this makes fast calculation of maximum likelihood estimates possible using generalized linear model techniques; in R idiom, the `glm()` function is used.

Novel analyses are presented on data drawn from the fields of social science and palaeoclimatology.

Acknowledgements

We thank Professor Gianfranco Lovison for helpful discussions; the AE and referees of the JSS review process provided constructive suggestions which improved the quality of the manuscript and associated software.

References

- Altham PME (1978). “Two Generalisations of the Binomial Distribution.” *Journal of the Royal Statistical Society C (Applied Statistics)*, **27**, 162–167.
- Altham PME, Hankin RKS (2012). “Multivariate Generalizations of the Multiplicative Binomial Distribution: Introducing the MM Package.” *Journal of Statistical Software*, **46**(12), 1–23. URL <https://www.jstatsoft.org/v46/i12/>.
- Brier SS (1980). “Analysis of Contingency Tables Under Cluster Sampling.” *Biometrika*, **67**(3), 591–596.
- Clisby KH, Sears PB (1955). “Palynology in Southern North America. Part III: Microfossil Profiles Under Mexico City Correlated with the Sedimentary Profiles.” *Bulletin of the Geological Society of America*, **66**(5), 511–520.
- Cox DR, Hinkley DV (1974). *Theoretical Statistics*. Chapman and Hall, Taylor and Francis Group Ltd, 2 Park Square, Milton Park, Abingdon, Oxford, OX14 4RN, UK.
- Danaher PJ, Hardie BGS (2005). “Bacon With Your Eggs? Applications of a New Bivariate Beta-Binomial Distribution.” *The American Statistician*, **59**(4), 282–286. ISSN 0003-1305. doi:10.1198/000313005X70939.
- Edwards AWF (1992). *Likelihood*. expanded edition. The Johns Hopkins University Press, 2715 North Charles Street, Baltimore, Maryland.
- Faegri K, Iversen J (1992). *Textbook of Pollen Analysis*. John Wiley & Sons, Hoboken, New Jersey.
- Foreman F (1955). “Palynology in Southern North America. Part II: Study of Two Cores from Lake Sediments of the Mexico City Basin.” *Bulletin of the Geological Society of America*, **66**(5), 475–510.
- Hankin RKS (2006). “Additive Integer Partitions in R.” *Journal of Statistical Software, Code Snippets*, **16**.
- Hankin RKS (2007). “Urn Sampling Without Replacement: Enumerative Combinatorics in R.” *Journal of Statistical Software, Code Snippets*, **17**(1).
- Johnson NL, Kemp AW, Kotz S (2005). *Univariate Discrete Distributions*. third edition. John Wiley & Sons, Hoboken, New Jersey.
- Lehmann EL (1966). “Some Concepts of Dependence.” *Annals of Mathematical Statistics*, **37**(5), 1137–1153.

- Lindsey JK, Mersch G (1992). “Fitting and Comparing Probability Distributions with log Linear Models.” *Computational Statistics & Data Analysis*, **13**(4), 373–384.
- Mosimann JE (1962). “On the Compound Multinomial Distribution, the Multivariate β -Distribution, and Correlations Among Proportions.” *Biometrika*, **49**, 65–82.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <https://www.R-project.org/>.
- Sears PB, Clisby KH (1955). “Palynology in Southern North America. Part IV: Pleistocene Climate in Mexico.” *Bulletin of the Geological Society of America*, **66**(5), 521–530.
- West LJ, Hankin RKS (2008). “Exact Tests for Two-Way Contingency Tables with Structural Zeros.” *Journal of Statistical Software*, **28**(11). URL <https://www.jstatsoft.org/v28/i11/>.
- Wilson JR (1989). “Chi-square tests for Overdispersion with Multiparameter Estimates.” *Journal of the Royal Statistical Society C (Applied Statistics)*, **38**(3), 441–453.

Affiliation:

P. M. E. Altham
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road Cambridge CB3 0WB
United Kingdom
E-mail: P.M.E.Altham@statslab.cam.ac.uk

Robin K. S. Hankin
University of Stirling
Scotland
E-mail: hankin.robin@gmail.com