Monte Carlo Approximation of Exact Conditional
Tests for Log-linear Models

James G. Booth
Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.
and
Ronald W. Butler
Department of Statistics, Colorado State University, Fort Collins, CO 80525, U.S.A.

September 1997

Summary

A simple but quite general simulation method for conducting exact conditional lack-of-fit tests in
log-linear models is proposed. Our Monte Carlo approximation utilizes an importance sampling method
motivated by the crude normal approximation to the Poisson distribution. Examples considered include
tests of quasi-symmetry and related models for square tables and tests concerning higher-order interac-
tions in multi-way tables. The method is competitive with direct simulation from the exact conditional
distribution when this is feasible and clearly outperforms alternative Monte Carlo procedures when di-
rect simulation is infeasible provided the degrees of freedom of the test is not too large. Extension of
the method to tests against non-saturated alternatives is straightforward and is briefly discussed and
illustrated.

1 Introduction

There is now an extensive literature on Monte Carlo methods for exact inference in contingency tables and,
more generally, in log-linear models. An excellent historical review is provided by Agresti (1992). More recent
research on this topic has focused on the use of Markov chain Monte Carlo (MCMC) techniques (Kolassa and
Tanner, 1994; Forster, McDonald and Smith, 1996; Smith, Forster and McDonald, 1996; Smith, McDonald,
Forster and Berrington, 1996), an idea originally proposed by Besag and Clifford (1989). In this paper we
show how exact inference can be carried out very efficiently for many standard models using a relatively
simple independent and identically distributed importance sampling technique. Applications we consider
include lack-of-fit tests for independence, quasi-independence, quasi-symmetry and uniform association in
two-way contingency tables and for higher-order interaction terms in multi-way tables. We emphasize exact
tests for lack-of-fit because it is for such tests that the standard chi-squared approximations often break
down. However, our methods also apply with some minor modifications to exact tests against non-saturated
alternatives. These modifications are illustrated for a test concerning a 3-factor interaction term in a 4-way
The methods we propose are motivated by the crude normal approximation to the Poisson distribution. Two importance distributions are given based on the resulting normal approximation to the conditional distribution of the data given the sufficient statistics. In both cases pseudo-datasets are obtained by rounding values generated from normal distributions in order that the importance distribution has the correct (integer) support. A naive approach is to simulate the pseudo-data values independently, ignoring the covariance terms in the multivariate normal approximate conditional distribution. However, this method is generally inefficient and can fail to converge even in some relatively simple problems. A second method captures the dependence structure by sequentially generating pseudo-data values as rounded normal variates conditionally on the previous rounded values. This method is shown to be extremely efficient for a wide range of standard models. In fact, the method appears to be competitive with direct simulation from the exact conditional distribution, when this is feasible; for example, in the test of independence in a two-way table. Moreover, provided the degrees of freedom of the test is not too large, efficient importance sampling clearly outperforms alternative Monte Carlo procedures that have been proposed for situations in which direct simulation is not feasible. In particular, in the problems we consider, our efficient importance sampling procedure is many times faster than the competing MCMC methods.

The rest of the paper is organized as follows. In Section 2 we briefly describe the sampling assumptions and log-linear model. Section 3 contains a discussion of exact conditional inference and Monte Carlo approximation of exact P-values. Section 4 concerns details of the normal approximation to the exact conditional distribution of the data given the sufficient statistics. Our naive and efficient importance distributions are described in Section 5. Formulas for the assessing Monte Carlo error are given in Section 6. Examples involving testing lack-of-fit are discussed in Section 7 and the extension to non-saturated models is given in Section 8. Finally, in Section 9 we discuss the limitations of our method.

2 Model and Notation

Suppose that data, \( y = (y_1, \ldots, y_n)^t \), arise from a Poisson log-linear model with means \( \mu = (\mu_1, \ldots, \mu_n)^t \) satisfying

\[
\log \mu_i = x_{i1} \beta,
\]

where \( x_i \) is a \( p \)-vector of known covariates and \( \beta \) is a \( p \)-vector of unknown regression parameters. The log-likelihood function for \( \beta \) is given by

\[
l(\mu; y) = \sum_{i=1}^n (y_i \log \mu_i - \mu_i).
\]
Let $\hat{\beta}$ denote the maximum likelihood estimate of $\beta$ and let $\hat{\mu}$ denote the corresponding vector of fitted values; i.e. $\hat{\mu}_i = \exp(x_i^T \hat{\beta})$, $i = 1, \ldots, n$. Two widely used measures of lack-of-fit for the model are the deviance,

$$D(y, \hat{\mu}) = 2 \{l(y, y) - l(\hat{\mu}, y)\},$$

and the Pearson chi-square statistic,

$$X^2(y, \hat{\mu}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$  

(4)

Standard asymptotic arguments imply these lack-of-fit statistics are approximately chi-squared variates with $n - p$ degrees of freedom provided the means are not too small.

In this paper we focus primarily on likelihood-ratio tests based on the deviance. However, the methods clearly extend to tests based on other measures of fit and test statistics.

3 Exact conditional inference

Let $s = X^T y$ denote the vector of sufficient statistics for $\beta$ in the Poisson log-linear model, where $X$ is the $n \times p$ covariate matrix. Then, the joint distribution of the data factors into

$$f(y; \mu) = f(y|s)f(s; \mu),$$

(5)

where the conditional distribution $f(y|s)$ does not depend on any unknown parameters. Thus, an "exact" test of lack-of-fit is obtained by considering the conditional distribution of $D(y, \hat{\mu})$ or $X^2(y, \hat{\mu})$ given $s$. Note here that the maximum likelihood estimate, $\hat{\beta}$, is a function of $s$ and hence so is the vector of fitted values, $\hat{\mu}$.

An exact conditional P-value for lack-of-fit based on the deviance statistic is defined as

$$P = P\{D(y, \hat{\mu}) \geq D_{obs}|s\} = \sum_{X'|y=s} 1_{\{D(x', \hat{\mu}) \geq D_{obs}\}} f(y|s).$$

In the majority of cases the normalizing constant for the conditional mass function, $f(y|s)$, is intractable. However, since the factorization in (5) implies that $f(y|s)$ is proportional to $f(y; \mu)$, we may write

$$P = \sum_{X'|y=s} 1_{\{D(x', \hat{\mu}) \geq D_{obs}\}} f(y; \mu)/\sum_{X'|y=s} f(y; \mu),$$

(6)

which does not depend on the particular value of $\mu$ satisfying (1). In particular, we may choose $\mu = \hat{\mu}$, the vector of fitted values. Thus, computation of $P$ involves evaluating $D(y, \hat{\mu})$ and $f(y; \hat{\mu})$ for all data configurations resulting in the observed value of $s$. Alternatively, if it is possible to simulate directly from
\[ P = \frac{1}{N} \sum_{k=1}^{N} 1\{ D_k^* \geq D_{\text{obs}} \} , \]  

where \( D_k^* = D(y_k^*, \hat{\mu}) \) is the value of the deviance computed using the \( k \)th simulated sample. Situations of this sort, in which exact inference is possible, are reviewed in Agresti (1992). However, there are many problems in which complete enumeration of (6) is computationally infeasible and simulation from the exact conditional distribution, as in (7), is not yet possible.

If the means are all large, the conditional distributions of \( D(y, \hat{\mu}) \) and \( X^2(y, \hat{\mu}) \) are approximately chi-squared with \( n - p \) degrees of freedom. That is, the conditional and unconditional approximations are the same. Thus, an approximation to \( P \) in (6) is given by

\[ P_X = P\left( \chi^2_{n-p} \geq D_{\text{obs}} \right) . \]

However, if some or all of the means are small, the chi-squared approximation can be very inaccurate and more exact methods are required. In addition, for some data configurations, conditioning on the sufficient statistics renders some cell entries redundant for inference about the model parameters. In such cases these cells should be removed from the data before applying the chi-squared approximation. For example, the quasi-symmetry model for the \( 8 \times 8 \) table considered by Smith, Forster and McDonald (1996) has \( p = 43 \) parameters and hence the degrees of freedom for the unconditional chi-squared approximation is 21. However, due to the sparseness of the table in this example, conditioning on the sufficient statistics places an additional 8 restrictions on the cell counts, reducing the degrees of freedom for the conditional chi-squared approximation to 13 (see Smith and McDonald (1995) for further discussion).

In this paper we describe a simple but quite general method for obtaining a Monte Carlo approximation to \( P \) via importance sampling. The method involves simulating pseudo-datasets from a distribution \( g(y|s) \) with the same support as \( f(y|s) \). An importance sampling approximation to the exact conditional \( P \)-value is then given by

\[ \bar{P} = \frac{1}{N} \sum_{k=1}^{N} \frac{f(y_k^*; \hat{\mu})}{g(y_k^*|s)} / \left( \sum_{k=1}^{N} f(y_k^*; \hat{\mu}) / g(y_k^*|s) \right) = \frac{1}{N} \sum_{k=1}^{N} 1\{ D_k^* \geq D_{\text{obs}} \} w_k / \sum_{k=1}^{N} w_k , \]

say, where \( w_k^* \) is the importance weight for the \( k \)th pseudo-dataset. A standard error for the importance sampling approximation can be obtained by the delta method. The number of importance samples, \( N \) in (9), can then be chosen to meet a predetermined absolute or relative accuracy criterion (see Section 6).

The rationale for approximation (9) is that the exact \( P \)-value satisfies the equation

\[ 0 = E_f \left( 1\{ D \geq D_{\text{obs}} \} - P \right) = E_g \left( 1\{ D \geq D_{\text{obs}} \} f(y|s) / g(y|s) - P f(y|s) / g(y|s) \right) , \]

\[ \text{Equation (10)} \]
where $E_f$ and $E_g$ denote expectation with respect to the conditional distributions $f(y|s)$ and $g(y|s)$ respectively. Hence, the factorization given in (5) and the Law of Large Numbers imply

$$\frac{1}{N} \sum_{k=1}^{N} \left( 1_{\{D_k \geq D_{res}\}} w_k^* - Pw_k^* \right) \to 0 \quad (11)$$

as $N \to \infty$ with probability one. The numerator of the $k$th importance weight in (9) is taken as

$$f(y_k^*; \hat{\mu}) = f(y_k^*|s)f(s; \hat{\mu}) \quad (12)$$

because the left side of (12) is readily determined, whereas individual terms of the right side are usually difficult to determine. The importance density is meant to mimic the first factor on the right side while the second factor cancels out in the ratio estimate given in (9).

4 Normal approximation

Our choice of importance distributions from which to simulate is motivated by the normal approximation to the Poisson distribution, $y_i \sim N(\mu_i, \mu_i)$. Since the counts, $y_1, \ldots, y_n$, are independent, their joint distribution is approximately multivariate normal with mean $\mu$ and covariance matrix $M$, where $M = \text{diag}(\mu)$.

Let $X_1$ and $X_2$ denote the matrices formed using the first $n - p$ rows and last $p$ rows of $X$ respectively, so that $X' = (X_1', X_2')$. Suppose also, without loss of generality, that the rows of $X$ have been arranged so that $X_2$ is full rank. Let $(y_1', y_2')'$ and $(\mu_1', \mu_2')'$ be the corresponding partitions of $y$ and $\mu$ respectively. Then,

$$\begin{pmatrix} y_1 \\ s \end{pmatrix} = \begin{pmatrix} I & 0 \\ X_1' & X_2' \end{pmatrix} y,$$

where $I$ is an identity matrix of rank $(n - p)$, and $0$ denotes a $(n - p) \times p$ matrix of zeros. It follows that the conditional distribution of $y_1$ given $s$ is approximately multivariate normal with mean

$$E(y_1|s) = E(y_1) - M_1 X_1 (X' M X)^{-1} \{ s - E(s) \}$$

and covariance matrix,

$$V(\mu) = M_1 - M_1 X_1 (X' M X)^{-1} X_1' M_1,$$  \quad (13)

where $M_1 = \text{diag}(\mu_1)$.

Recall that the true conditional distribution of $y_1$ given $s$ does not depend on $\beta$. Thus, we are free to choose any value of $\beta$ when constructing an approximate conditional distribution. In particular, the choice, $\beta = \hat{\beta}$, implies $\mu = \hat{\mu}$ and hence $E(s) = X' \mu = X' \hat{\mu} = s$. Thus, with this choice of $\beta$ the formula for the conditional mean in the normal approximation reduces to

$$E(y_1|s) = (I \ 0) \hat{\mu} = \hat{\mu}_1$$  \quad (14)
and the conditional variance becomes $V(\mu)$.

5 Importance Distributions

Note that the normal approximation derived in the previous section is not a mass function with the same support as $f(y|s)$ and therefore is not suitable as an importance distribution for use in (9). One possibility is to generate vectors from the approximating normal distribution and then to round each element to the nearest integer. The problem with this approach is that calculation of the probability mass associated with the resulting pseudo-dataset involves an intractable $n-p-1$ dimensional integral.

5.1 A Naive Importance Distribution

One way to circumvent the intractable integration problem is to ignore the covariances in the multivariate normal approximating distribution. A naive importance sampling algorithm then proceeds as follows:

Step 1: Generate a random vector from the multivariate normal distribution with mean $\hat{\mu}$, and variance matrix $\text{diag}(\hat{v}_i)$, where $\hat{v}_{ij}$ is the $(i, j)$th element of $V(\hat{\mu})$ given in (13). The pseudo-data $y_i = (y_{i1}, \ldots, y_{i,n-p})^T$ then consists of the elements of the multivariate normal vector rounded to the nearest integer.

Step 2: Solve for the remaining pseudo-data values, $\hat{y}_2 = (y_{n-p+1}, \ldots, y_n)^T$, using the linear equations, $X_k^T y^* = s$, or equivalently

$$\sum_{i=n-p+1}^{n} x_{ij} \hat{y}_i = s_j - \sum_{i=1}^{n-p} x_{ij} \hat{y}_i, \quad j = 1, \ldots, p.$$ 

Note that this system of equations has a unique solution because the $p \times p$ matrix, $X_2$, is full rank.

Step 3: Check that all pseudo-data values are non-negative. If any values are negative return to Step 1.

(Acceptance rates for pseudo-datasets generated in several examples are recorded in Table 2.)

Step 4: Calculate the probability mass associated with the pseudo-data,

$$g(y^*|s) \propto \prod_{i=1}^{n-p} \left\{ \Phi \left( \frac{y^*_i + 0.5 - \hat{\mu}_i}{\sqrt{v_{ii}}} \right) - \Phi \left( \frac{y^*_i - 0.5 - \hat{\mu}_i}{\sqrt{v_{ii}}} \right) \right\}.$$

Step 5: Calculate the model deviance, $D^* = D(y^*, \hat{\mu})$, (or other lack-of-fit statistic) based on the pseudo-data.

Note that the fitted values, $\hat{\mu}$, are the same for the pseudo-data because the sufficient statistics are unchanged.

Step 6: Repeat Steps 1-5 until the importance sampling approximation to the exact P-value given in (9) converges (see Section 6).
This importance sampling method is attractive because of its simplicity. However, in practice ignoring
the dependence between the counts due to conditioning results in an extremely inefficient Monte Carlo
approximation. In fact, in two examples considered in Section 7, involving large sparse tables, we were
unable to obtain convergence even after generating several million pseudo-datasets. We now describe an
alternative way of constructing pseudo-datasets which takes into account much of the dependence between
the counts resulting in a dramatic improvement in computational efficiency.

5.2 An Efficient Importance Distribution

Let \( z = (z_1, \ldots, z_{n-p})^t \) denote a multivariate normal variable with mean vector \( \mu \) and covariance matrix
\( V(\mu) \). That is, \( z \) has the approximating normal distribution for \( f(y|s) \) derived in Section 4. As noted above,
rounding the components of simulated \( z \) vectors does not lead to a tractable importance distribution. Our
efficient importance sampling algorithm is an attempt to achieve this goal approximately by replacing Steps
1 and 4 in the previous algorithm by

Step 1: Generate \( z_1 = z_1^* \) from its marginal normal distribution with mean \( m_1 = \mu_1 \) and variance \( s_{11} = \sigma_{11} \)
and define \( y_i^* \) to be the value of \( z_i^* \) rounded to the nearest integer. For \( k = 1, \ldots, n - p - 1 \) define \( y_{k+1}^* \)
to be the rounded value of \( z_{k+1}^* \), where \( z_{k+1}^* \) is drawn from the normal distribution with mean
\[ m_{k+1} = E(z_{k+1}^*|z_i^* = y_i^*, i = 1, \ldots, k) \] (15)
and variance
\[ s_{k+1,k+1} = \text{Var}(z_{k+1}^*|z_i^* = y_i^*, i = 1, \ldots, k). \] (16)

The standard matrix formulas for the conditional mean and variance in (15) and (16) can be found, for
example, in Mardia, Kent and Bibby (1979, Section 3.2). It is, however, more efficient computationally
to obtain these conditional moments sequentially using univariate calculations that do not involve
matrix inversion. Details of such calculations are provided in the Appendix.

Step 4: Calculate the probability mass associated with the pseudo-dataset,
\[ g(y^*|s) \propto \prod_{i=1}^{n-p} \left\{ \Phi \left( \frac{y_i^* + .5 - m_i}{\sqrt{s_{ii}}} \right) - \Phi \left( \frac{y_i^* - .5 - m_i}{\sqrt{s_{ii}}} \right) \right\}. \]

A possible modification of both importance distributions is to replace each univariate normal distribution
with a \( t \)-distribution after a location and scale transformation to obtain the desired mean and variance. In
some contexts use of the heavier tailed \( t \)-distribution helps to stabilize the importance weights and speeds up
the convergence of the approximation. However, in the context considered in this paper replacing normals
Table 1: The number of simulated datasets required to attain a 5\% relative error criterion (RE = .05) with 99\% probability (\(\alpha = .01\)). (\(P = \) true P-value, \(N = (z_{\alpha/2}/RE)^2(1 - P)/P\))

with t-variates does not appear to help. A possible explanation is that conditioning bounds the support of the data and hence heavy tails are unnecessary.

6 Monte Carlo Error

Let \(u = w1(D \geq D_{obs})\). Then the Monte Carlo error in (9) can be assessed using the asymptotic variance formula derived using the delta method,

\[
\text{Var}_g(\hat{P}) \approx \frac{1}{N} E_g(w)^2 E_g\{(u - wP)^2\} = \frac{\sigma^2}{N},
\]
say. An approximation to \(\sigma\) based on the \(N\) simulations is

\[
\hat{\sigma} = \frac{1}{\bar{w}} \sqrt{\frac{1}{N} \sum_{k=1}^{N} (w_k - \bar{w} \cdot \hat{P})^2}.
\]

To control the Monte Carlo error in practice we can continue the simulations until the absolute error is estimated to be less than a prespecified level, \(\epsilon\), with 100(1 - \(\alpha\))\% confidence; i.e. until

\[
\text{AE} = z_{\alpha/2} \hat{\sigma}/\sqrt{N} \leq \epsilon, \tag{17}
\]

where \(z_{\alpha}\) denotes the \((1 - \alpha)\)-quantile of the standard normal distribution. Alternatively, we can stop the simulations based on the relative error criterion

\[
\text{RE} = \frac{\text{AE}}{\hat{P}} \leq \epsilon. \tag{18}
\]

In the examples discussed in the next two sections we used a 5\% relative error criterion (\(\epsilon = .05\)) and a 99\% confidence level (\(\alpha = .01\)). However, in practice, if the P-value is very small then such accuracy is typically more than is necessary. One possibility is to switch to an absolute error criterion of \(\epsilon = .0005\) if \(\hat{P} < .01\).

In the next section we give actual computing times for our method in several examples. While it is interesting to know whether a technique takes seconds, minutes or hours, we feel that these numbers are somewhat misleading because they depend so much on the computer, its compiler, the programming language used, and last but not least the skill of the programmer. Although different Monte Carlo approximation methods require different amounts of computational overhead (such as the calculation of importance weights) a more accurate assessment of computational efficiency is generally provided by the number of simulated
<table>
<thead>
<tr>
<th>Example</th>
<th>Method</th>
<th>Deviance</th>
<th>df</th>
<th>$P_X$</th>
<th>$\bar{P}$</th>
<th>$N \times 10^{-3}$</th>
<th>Accept%</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Poisson distribution</td>
<td>Efficient</td>
<td>25.25</td>
<td>19</td>
<td>.1524</td>
<td>.1859</td>
<td>25</td>
<td>94</td>
<td>00:29</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.1924</td>
<td>63</td>
<td>68</td>
<td>00:44</td>
</tr>
<tr>
<td>2. Independence in a 4 x 4 table</td>
<td>Efficient</td>
<td>15.49</td>
<td>9</td>
<td>.0784</td>
<td>.1091</td>
<td>26</td>
<td>96</td>
<td>00:14</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.1120</td>
<td>443</td>
<td>93</td>
<td>03:28</td>
</tr>
<tr>
<td>3. Complete independence in a 2 x 2 x 3 table</td>
<td>Efficient</td>
<td>14.93</td>
<td>7</td>
<td>.0370</td>
<td>.0674</td>
<td>56</td>
<td>88</td>
<td>00:22</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.0660</td>
<td>1944</td>
<td>26</td>
<td>19:53</td>
</tr>
<tr>
<td>4. Higher-order interactions in a 2^4 table</td>
<td>Efficient</td>
<td>6.26</td>
<td>4</td>
<td>.1803</td>
<td>.2164</td>
<td>13</td>
<td>89</td>
<td>00:08</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.2199</td>
<td>20</td>
<td>89</td>
<td>00:12</td>
</tr>
<tr>
<td>5. Quasi-symmetry in a 8 x 8 table</td>
<td>Efficient</td>
<td>19.12</td>
<td>13</td>
<td>.1194</td>
<td>.0637</td>
<td>60</td>
<td>68</td>
<td>03:45</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.2199</td>
<td>20</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>6a. Quasi-independence in a 5 x 5 table</td>
<td>Efficient</td>
<td>13.55</td>
<td>5</td>
<td>.0187</td>
<td>.0217</td>
<td>253</td>
<td>82</td>
<td>01:43</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.0229</td>
<td>2685</td>
<td>41</td>
<td>27:12</td>
</tr>
<tr>
<td>6b. Uniform association in a 5 x 5 table</td>
<td>Efficient</td>
<td>16.21</td>
<td>13</td>
<td>.2377</td>
<td>.0402</td>
<td>261</td>
<td>74</td>
<td>04:53</td>
</tr>
<tr>
<td></td>
<td>Naive</td>
<td></td>
<td></td>
<td></td>
<td>.08</td>
<td></td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Monte Carlo approximations of exact conditional $P$-values obtained by importance sampling using 5\% relative error criterion and a 99\% confidence level. ($N$ = number of pseudo-datasets used, $A$ = acceptance rate for pseudo-datasets; thus, $100N/A$ = number of pseudo-datasets generated)

datasets that are required to reach fixed convergence criteria. A good reference point, therefore, is the number of samples required under ideal circumstances; that is, when direct simulation from the exact conditional distribution is possible and approximation (7) is used. These values are given in Table 1 for a selection of values of $P$.

7 Examples

Table 2 summarizes the results of applying the importance sampling method once in each of six examples that have appeared in the literature. All of the computations were done on a SUN Sparc10 computer using Fortran. Also, to ensure honesty in our calculations we used the same seed in all our examples.

Example 1 concerns the test of the Poisson distributional assumption for the number of deaths by horse-kick in ten Prussian cavalry corps over a twenty year period 1875-1895 (see Preece, Ross and Kirby, 1988). Complete enumeration is straightforward in this case because the exact conditional distribution is symmetric multinomial. Note also that comparison with the numbers in Table 1 suggests that Monte Carlo approximation via simulation from the exact distribution would require roughly half the number required by our efficient importance sampling method.

Example 2 concerns the test of independence in a 4 x 4 table for data on ratings of sexual enjoyment by husbands and wives (Hout, Duncan and Sobel, 1987; Agresti, 1990; Kolassa and Tanner, 1994). This problem is interesting because it is a situation in which complete enumeration is possible using the StatXact
software package (StatXact, 1991). In this instance it took StatXact 13 minutes and 19 seconds to obtain the exact P-value, \( P = .1137 \). This time compares with 3 minutes and 28 seconds to obtain an accurate Monte Carlo approximation using our naive importance distribution and only 14 seconds using our efficient importance distribution. The test of independence in a two-way contingency table is also a situation in which direct simulation from the exact conditional distribution is possible (Agresti, Wackerly and Boyett, 1979; Patefield, 1981). Monte Carlo approximation of \( P \) using this approach takes StatXact about 15 seconds and 20,000 samples to achieve our accuracy requirements. Note that the number of samples is in fairly good agreement with the number predicted in Table 1 when \( P = .1 \). Our importance sampling method required 25,000 samples in this example illustrating remarkable efficiency relative to direct simulation.

Kolassa and Tanner (1994) consider exact P-value calculation for the hypothesis of complete independence in a \( 2 \times 2 \times 3 \) table using MCMC methods utilizing a saddlepoint c.d.f. approximation for the hypergeometric distribution. The original data were reported by Muller and Mayhill (1971) (see also Bishop, Fienberg and Holland, 1975). Kolassa and Tanner report a P-value of \( P = .063 \) based on 5000 independent Markov chains of length 200; that is, 1 million dependent simulated datasets. It is difficult to compare the efficiency of their method with the importance sampling methods because no error bounds are reported in their paper. However, we note that using our efficient importance distribution we required only 56 thousand independent pseudo-datasets to achieve our 5% relative error bound with 99% confidence.

Example 4 concerns the lack of fit of a model for a \( 2^4 \) table reported by Dykes, Lazzara, Ahmann, Blumenstein, Schwartz and Brann (1980) and later analysed by Morgan and Blumenstein (1991). Letting \( I, H, V \) and \( R \) denote the four binary factors, the lack-of-fit test concerns the model \( HRV+HI+RI+VI \). Note that the computations are extremely fast using either importance distribution. This example is discussed further in the next section.

Smith, Forster and McDonald (1996) consider exact P-value calculation for testing quasi-symmetry in an \( 8 \times 8 \) table on husband’s and wife’s ethnicity using the Hastings-Metropolis algorithm. They report a P-value of \( P = .0602 \) and a 99% confidence interval of \( (.0546, .0658) \) based on a Markov Chain of length 1 million. Thus, their method achieved a relative error bound of about 10%, suggesting that close to 4 million simulated samples would be required to achieve our 5% bound. This can be contrasted with our importance sampling approximation which required 59 thousand pseudo-datasets to converge. These numbers suggest that importance sampling is at least an order of magnitude more efficient than competing MCMC methods in this problem. The acceptance rate for pseudo-datasets was only about 0.3% using the naive importance distribution. This was due to the sparseness of much of the contingency table and represents an extreme example of the need to incorporate the dependence structure in the importance distribution. In fact, we
were unable to achieve our convergence criteria using the naive method.

Example 6 concerns tests of quasi-independence and uniform association in a 5 \times 5 table of two different pathologists ratings of carcinoma (Agresti, 1988; Holmquist, McMahon and Williams, 1987). McDonald, DeRoure and Michaelides (1995) calculate the exact P-value for testing lack-of-fit of the quasi-independence model by enumerating all possible tables with the same sufficient statistics reporting the value \( P = .0225 \). In another paper Smith and McDonald (1995) use a simulate and reject algorithm to obtain a Monte Carlo approximation to the exact P-value in this problem. Their method involves simulating tables with fixed margins under the assumption of independence after replacing the diagonal counts with values as close as possible to the expected values under independence. Simulated tables are accepted if their diagonal values agree with the replacement values. Smith and McDonald report an approximate P-value with a relative error of about 10\% with 99\% confidence based on 30,000 accepted tables and an acceptance rate of about 6\%. Thus, using their method, approximately 2 million tables would need to be generated in order to accept the 120,000 pseudo-tables required to achieve our 5\% relative error target. In contrast, our importance sampling method required a total of 308,000 tables with an acceptance rate of 82\% to achieve the same level of accuracy. This comparison suggests that, although our method incurs some loss of efficiency through the use of an inexact importance distribution, this is offset substantially by the higher acceptance rate of simulated tables.

McDonald et al. (1995) report an approximate conditional P-value of .0370 for the uniform association model based on the deviance which they obtained using the MCMC method of Forster et al. (1996) and Smith, Forster and McDonald (1996) cited earlier. The accuracy of their approximation and the length of the Markov Chain used are not reported. However, we note that their approximation is not within the 5\% relative error bounds of the value we obtained using importance sampling. Note that the naive importance sampling method failed to converge in this example.

8 Non-saturated Alternatives

In many problems involving contingency table data it is of interest to compare a null model with a non-saturated but more general alternative log-linear model. Two common examples are testing marginal homogeneity in a square table (e.g. symmetry versus quasi-symmetry) and testing conditional independence versus a common odds-ratio in a \( 2 \times 2 \times k \) table. The methods described in the previous sections can easily be extended to such cases as follows.

Let \((s, t)\) denote the vector of sufficient statistics under the alternative model, where \( s = X^t y \) as defined previously. Let \( G(s, t) \) denote a generic test statistic for testing the fit of the null model versus the alternative.
We will suppose that large values of $G$ lead to rejection of the null model. Then, an exact conditional P-value for the test is given by

$$P = P\{G(s, t) \geq G_{obs} | s\} = \sum_{x^*y=s} 1_{\{G(x^*, t) \geq G_{obs}\}} f(y|s).$$  \hspace{1cm} (19)$$

A Monte Carlo approximation to the exact P-value in (19) is given by

$$\hat{P} = \frac{\sum_{k=1}^{N} 1_{\{G_k^* \geq G_{obs}\}} \frac{f(y_k^*; \hat{\mu})}{\hat{g}(y_k^*|s)}}{\sum_{k=1}^{N} \frac{f(y_k^*; \hat{\mu})}{\hat{g}(y_k^*|s)}} = \frac{\sum_{k=1}^{N} 1_{\{G_k^* \geq G_{obs}\}} w_k^*}{\sum_{k=1}^{N} w_k^*},$$ \hspace{1cm} (20)

where $G_k^* = G(s, t_k^*)$ is the value of $G$ obtained using the $k$th pseudo-dataset generated from the importance distribution $g(y|s)$. Note that the computations required in (20) are essentially identical to those required to evaluate the P-value for lack-of-fit given in (9) provided that $G$ is an explicit function of the sufficient statistics $(s, t)$ and the maximum likelihood estimates for the null model. This is the case, for example, if $G$ is the efficient score statistic (see McCullagh and Nelder, 1989, Section 12.3) but not for the likelihood-ratio test. In general, computation of (20) with $G$ equal to the likelihood-ratio statistic involves refitting the alternative model to each pseudo-dataset.

To illustrate the extension to non-saturated alternatives we revisit the $2^4$ table analyzed by Morgan and Blumenstein (1991). These authors considered the model $HRV + HIR + IV$ as an alternative to $HRV + HI + RI + VI$ discussed in Section 7. This alternative model has a single additional parameter corresponding to the $HIR$ three-factor interaction term. A one-sided significance test can therefore be based on the statistic $G = t$ or $G = -t$, where $t$ is the univariate sufficient statistic corresponding to the additional parameter. For this example the observed value of $t$ is in the left tail of the conditional distribution and hence we use $G(t) = -t$ to determine an exact P-value. Using our efficient importance sampling method and the same convergence criteria as in Section 7 (i.e. RE = .05 and $\alpha = .01$), we obtained an approximate P-value of $\hat{P} = .0281$ using 112 thousand pseudo-datasets. The computing time was 72 seconds. The exact P-value in this example is reported by Morgan and Blumenstein (1991) as $P = .030$ which is slightly outside our 5% error bounds. A second run using a different seed resulted in the value $\hat{P} = .0292$. For the same problem Forster et al. (1996) report an approximate P-value of .029 with a 99% confidence interval (.026,.032) based on a Gibbs chain of length 100 thousand. It follows that it would require a chain of length around 400 thousand to attain our 5% relative error criterion using their method.

9 Discussion

We have described a simple method for conducting Monte Carlo exact conditional tests in log-linear models via importance sampling. In the examples presented, our method is competitive with direct simulation from
the exact conditional distribution when this is possible and clearly outperforms competing Monte Carlo methods when direct simulation is not possible. In particular, importance sampling appears to be many times faster than competing MCMC methods. In this regard, it is interesting to note that, whereas our method is roughly forty times more efficient than the Hastings-Metropolis Markov chain method in the quasi-symmetry example presented in Section 7, it is only about 4 times faster than the Gibbs sampling Markov chain method for the example presented in Section 8. This difference could merely be due to the specific nature of the examples presented. However, an alternative explanation is that the standard errors for the two different MCMC approximations were computed using completely different methods. Monte Carlo error for the Hastings-Metropolis approximation of Smith, Forster and McDonald (1996) was calculated using the window variance estimate described by Geyer (1992). In contrast, the error in the Gibbs approximation of Forster et al. (1996) was calculated using the method of Raftery and Lewis (1992) which is known to be quite sensitive to the initial behavior of the chain (Cowles and Carlin, 1996).

We conclude by noting that in the examples discussed in the previous two sections the degrees of freedom for testing lack-of-fit of the null model is never extremely large, ranging from 1 to 19. This is typical in many practical situations involving contingency table data. In such cases our (efficient) importance sampling method always works extremely well. However, the efficiency of importance sampling decreases as the dimension of the importance distribution increases. Thus, for problems involving extremely large tables the methods presented in this paper may not be useful. For example, Whittaker (1990, Section 9.3) fits a model to a sparse 2^8 table in which all 3-factor and higher-order interactions are assumed to be zero. In this case the number of cells is n = 256 and p = 37 leaving 219 degrees of freedom for testing lack-of-fit. In such problems our importance sampling method completely breaks down. In fact, more than 99.9% of pseudo-samples are rejected because of negative counts. In contrast MCMC methods may still be applied in such settings (Forster et al., 1996) and hence are likely to remain the method of choice for high dimensional problems.

Acknowledgment

The authors are grateful for some helpful comments from Alan Agresti and Jim Hobert.

Appendix

Let $z_k^*$, $m_k$, $s_{kk}$, $\mu_k$ and $\nu_{ki}$ be defined as in Section 5 and let $x_k = z_k^* - m_k$, $k = 1, \ldots, n-p$. It follows from their construction that the $x_k$'s are independent normal variates with zero means and variances $s_{kk}$. Let $\nu'_{ki} =$
\[ \text{Cov}(\mathbf{x}_k, \mathbf{z}_i), \text{ for } 1 \leq k < i \leq n - p. \] Then the covariance matrix of the vector \((x_1, \ldots, x_{k-1}, x_k^*, \ldots, x_{n-p}^*)^t\) has the form

\[
\begin{pmatrix}
    s_{11} & v_{1k}^* & \cdots & v_{1,n-p}^*
    \\
    \vdots & \ddots & \vdots & \vdots \\
    v_{k1}^* & \cdots & s_{k-1,k-1} & v_{k-1,k}^* & \cdots & v_{k-1,n-p}^*
    \\
    \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
    v_{n-p,1}^* & \cdots & v_{n-p,k}^* & \cdots & s_{n-p-1,n-p-1} & v_{n-p,n-p}^*
\end{pmatrix}
\]

Standard results for multivariate normal vectors (for example, Mardia et al., 1979, Section 3.2) imply that, for \(k = 2, \ldots, n - p,\)

\[
m_k = E(x_k^* | x_1, \ldots, x_{k-1}) = \mu_k - \sum_{i=1}^{k-1} \frac{x_i v_{ik}^*}{s_{ii}}
\]

and

\[
s_{kk} = \text{Var}(x_k^* | x_1, \ldots, x_{k-1}) = s_{kk} - \sum_{i=1}^{k-1} \frac{(v_{ik}^*)^2}{s_{ii}}.
\]

Finally, note that \(v_{1,k+1}^* = v_{1,k+1} \) and, for \(i = 2, \ldots, k,\)

\[
v_{i,k+1}^* = v_{i,k+1} - \sum_{j=1}^{i-1} \frac{v_{ij}^* v_{j,k+1}}{s_{jj}}.
\]

Hence, the covariance terms required in the calculation of (15) and (16) can also be obtained using sequential univariate computations.

References


