MCMC in I x J x K Contingency Tables

Florentina Bunea Julian Besag

NRCSE

Technical Report Series

NRCSE-TRS No. 037

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

Fields Institute Communications
Volume **00**, 0000

MCMC in $I \times J \times K$ contingency tables

Florentina Bunea
University of Washington
fbunea@stat.washington.edu

Julian Besag
University of Washington
julian@stat.washington.edu

Abstract. The paper reviews Markov chain Monte Carlo exact tests for assessing the goodness of fit of probability models to observed datasets. All unknown parameter values are removed from the analysis by the standard device of conditioning on sufficient statistics, which in turn leads to constrained target distributions. It is easy to design algorithms that maintain these distributions but often very difficult to ensure irreducibility. Strangely, this condition is not required for validity of the p–value; nevertheless it is generally desirable. One of the simplest taxing problems occurs in testing for no three–way interaction in $2 \times J \times K$ contingency tables and in this case the paper provides a corresponding irreducible chain. Although appropriate modifications of the algorithm can be used to test for any hierarchical model in quite general multidimensional contingency tables, irreducibility is no longer guaranteed. Thus, the paper identifies a class of easily posed problems of interest to statisticians, where the construction of irreducible chains presents a challenging task.

1 Introduction

A problem of widespread interest to statisticians concerns the goodness of fit of probability models to observed datasets. The assessment of models for multidimensional contingency tables provides a common example and the particular context of the present paper. Often such tables are too sparse to apply standard, typically asymptotic, distribution theory and it is necessary to devise alternative methods of assessment. The complications that can arise are exemplified most easily in testing for no three–way interaction in three–dimensional contingency tables; that is, $I \times J \times K$ tables x, formed by layers $i = 1, \ldots, I$, rows $j = 1, \ldots, J$ and columns $k = 1, \ldots, K$, in which each cell (i, j, k) contains a count x_{ijk}, corresponding to the frequency of a particular event indexed by the triple (i, j, k). In fact, we shall focus especially on the case $I = 2$, which is important in *logistic regression*; see Cox and Snell [**9**], for example. Our approach, which provides an exact p–value via Markov chain Monte Carlo (MCMC) simulation, applies to arbitrary I and indeed to the assessment of any hierarchical model in higher–dimensional tables but irreducibility of the algorithm may fail. We return to this point in Section 4, where we identify a corresponding class of open problems in devising MCMC algorithms for constrained distributions.

There are several formulations of multidimensional contingency tables, all of which are equivalent in the context of the paper. For definiteness, we adopt that in which the x_{ijk}'s are generated according to a multinomial distribution with unknown cell probabilities p_{ijk} and pre–specified sample size x_{+++}, where + denotes summation over the corresponding index and $p_{+++} = 1$. Thus, the table x has probability

$$\frac{x_{+++}!}{\prod_{i,j,k} x_{ijk}!} \prod_{i,j,k} p_{ijk}^{x_{ijk}} . \tag{1.1}$$

Then it is often of interest, especially at an initial stage of statistical analysis, to investigate whether an observed table $x^{(1)}$ is compatible with a particular parameterization of the p_{ijk}'s. Two examples follow.

Example 1. A common hypothesis is that the layer, row and column categorizations of the table are independent, which implies that $p_{ijk} = \theta_i \phi_j \psi_k$, where the θ_i's, ϕ_j's and ψ_k's form arbitrary probability distributions. Exact tests for independence in multidimensional contingency tables are generally amenable to the simple Monte Carlo approach described in Section 2.

Example 2. The least restrictive non–saturated model is obtained by allowing two–way but not three–way interaction among the p_{ijk}'s, as in Bartlett [**2**]. That is,

$$p_{ijk} \;=\; \theta_{ij}\, \phi_{ik}\, \psi_{jk} \,, \tag{1.2}$$

in which the unknown θ_{ij}, ϕ_{ik} and ψ_{jk} ensure that the p_{ijk}'s form a valid probability distribution. Testing for the absence of three–way interaction in a three–way table is the simplest non–trivial task in analyzing multidimensional contingency tables.

Any such hypothesis, also in higher–dimensional tables and in more general settings, can be tested in principle by conditioning on sufficient statistics for the unknown parameters and then assessing the plausibility of $x^{(1)}$ as a random draw from the resulting, now completely determined, conditional distribution $\{\pi(x) : x \in S\}$, where S denotes the constrained sample space. We say more about this in Section 2 but first we return to the above examples and identify the corresponding reference distributions π.

Example 1. In testing for independence, the sufficient statistics are the one–dimensional layer, row and column margins, $x^{(1)}_{i++}$, $x^{(1)}_{+j+}$ and $x^{(1)}_{++k}$ for each i, j and k, so that S is the set of all contingency tables with these totals and

$$\pi(x) \;\propto\; 1 \,/ \prod_{i,j,k} x_{ijk}! \,, \qquad\qquad x \in S \,. \tag{1.3}$$

Example 2. In testing for no three–way interaction, the sufficient statistics are all margins of the form $x^{(1)}_{ij+}$, $x^{(1)}_{i+k}$ and $x^{(1)}_{+jk}$ and, if we condition on these, $\pi(x)$ is again given by (1.3) but now S is the more restrictive set of all contingency tables with the same two–dimensional margins as $x^{(1)}$.

The remainder of the paper is set out as follows. In Section 2, we review three methods of calculating exact p–values for conditional tests. In Section 3, we describe an irreducible MCMC algorithm to solve the problem in Example 2 with $I = 2$, drawing on Besag and Clifford [**5**] and Diaconis and Sturmfels [**10**]. Section 4 discusses some other related work, extensions of the MCMC approach to higher–dimensional tables and some corresponding open problems. Proofs of results in Section 3 are given in an appendix.

2 Construction of exact p–values

2.1 Traditional p–values. Suppose that we wish to assess the compatibility of data $x^{(1)}$ with a fully specified probability model $\{\pi(x) : x \in S\}$. Some informal tools are available but the most common (frequentist) procedure is to carry out a *significance test* and report a corresponding p–value, based on some test statistic $u : S \to R$, with the property that an unusually large (say) observed value $u^{(1)}$ of u suggests that the model is false. More precisely, suppose that $u^{(1)}$ lies at the upper $100\alpha\%$ point of the distribution of u induced by π. Then α is declared to be the p–value for the test. What is meant here is that, if in fact the model is correct, we would obtain a value of u as or more extreme than the observed $u^{(1)}$ on a proportion α of occasions. In particular, if α is small, there is evidence of a conflict between the data $x^{(1)}$ and the proposed model.

We emphasize that the p–value for an observed test statistic $u^{(1)}$ is calculated merely from the postulated distribution π but nevertheless it is important to have a plausible alternative model in mind and to choose a u that is sensitive to the differences between the competing hypotheses. Often a generalized likelihood ratio test statistic provides a good choice, though here we do not wish to rely on asymptotics. We also note that a p–value is sometimes misinterpreted as representing the probability that the proposed model is correct. In fact, it is evident that, in practice, almost *any* statistical model is false, at least to some degree. At first sight, it may seem that this undermines the whole idea of a p–value, a point we discuss below.

Thus, it is often claimed that failure of a test to "reject" a proposed model merely reflects a lack of sufficient data or a poor choice of test statistic. Although this argument has a superficial appeal, it ignores an important purpose of statistical modelling, namely to provide a parsimonious representation (ideally, of course, an explanation) of the available dataset and of others that might have been obtained in its stead. If there is no clear conflict between the proposed model and the observed data, then a more complex formulation is unlikely to be warranted and may even be detrimental, though this by no means implies that modifications would not be required to represent a more extensive dataset. Fisher [**11**], page 314 states: "More or less elaborate forms will be suitable according to the volume of the data". Thus, the outcome of a statistical test refers to the compatibility between the model and the data and not to some grander view of the model itself. Note here that, in exploratory analysis, we might carry out several different tests without necessarily taking explicit account of the effects of "multiple testing", though the results of all tests should be reported. If a particular test statistic suggests an interesting conflict between the data and the model, then it does so regardless of how many other tests have been performed. Incidentally, the above remarks explain why we view p–values as useful at an initial rather than at a confirmatory stage of data analysis.

As regards the assessment of probability models for contingency tables, the enumeration of S is rarely feasible, except for rather small datasets, and this usually prevents the direct calculation of α from π. Also, the usual chi–squared approximation for the distribution of the deviance and for the closely related Pearson's X^2 statistic often breaks down in three– and higher–dimensional tables because some of the expected counts are too small; and approximations for other test statistics do not generally exist.

2.2 Barnard's Monte Carlo exact p–values. A different general approach, that does not require complete enumeration or asymptotic approximation, is based on simulation and was first proposed by Barnard [**1**]. In the context of contingency tables, again let $u^{(1)}$ denote the value of a particular test statistic u for an observed table $x^{(1)}$ and let $u^{(2)}, \dots, u^{(n)}$ be the corresponding values for a random sample of $n-1$ tables $x^{(2)}, \dots, x^{(n)}$ generated from the reference distribution π. Then, if $x^{(1)}$ is also from π, the n values of u form a random sample from a single distribution and, ignoring the possibility of ties, the rank of $u^{(1)}$ among all n values is drawn from a uniform distribution on the integers $1, \dots, n$. Thus, if large values of $u^{(1)}$ suggest a conflict with π and if, in the event, $u^{(1)}$ ranks qth largest among $u^{(1)}, \dots, u^{(n)}$, we may declare an exact p–value q/n. When ties with $u^{(1)}$ occur, we recommend reporting the corresponding range of p–values, though a more rigorous solution is also available, as in Besag and Clifford [**5**], for example. We ignore ties in the remainder of the paper. As regards the size of the random sample, $n = 1000$ generally gives very close agreement between different individuals, though we often prefer $n = 10000$. If there are computational constraints, one can also the more frugal *sequential* version of the test in Besag and Clifford [**6**]. Note that Barnard's approach is distinct from, though asymptotically equivalent to, the more obvious procedure in which a random sample from π is used to construct an empirical distribution function for u and a corresponding approximate p–value. The distinction becomes more important in devising MCMC versions, as in Section 2.3.

For applications of Barnard's Monte Carlo procedure, see, for example, Besag and Diggle [**7**], Besag [**4**] and Guo and Thompson [**12**]. Unfortunately, for most models of interest in multidimensional contingency tables, methods of drawing random or, more correctly, pseudo–random samples from the salient reference distribution π have not been devised. Testing for independence, as in Example 1 or correspondingly in higher–dimensional contingency tables, provides a rare exception; see e.g. Patefield [**15**]. When simple Monte Carlo tests are not available, as in Example 2, we can turn instead to MCMC procedures, as described below.

2.3 MCMC exact p–values. As before, let $\{\pi(x) : x \in S\}$ denote the target distribution and now suppose that P is a Markov transition probability matrix with state space S and stationary distribution π. In using P to simulate a Markov chain and create samples from which an exact p–value can be calculated, we encounter three different problems: burn–in, dependence between samples, and irreducibility, each of which has a feature that is unusual in MCMC. Thus, the first problem is trivial to avoid, merely by initiating the chain with the observed data $x^{(1)}$. This is legitimate because, as emphasized in Section 2.1, the calculation of a p–value is based on the supposition that $x^{(1)}$ is indeed drawn from π.

The second problem is more irksome because we cannot appeal to the ergodic theorem, as we do in calculating expectations, and it is against the spirit of MCMC and may not be feasible to leave long enough gaps to ensure virtual independence between successive samples. Initiation of the chain by $x^{(1)}$ ensures that, in forming a rationale for the p–value, we can assume that all samples are drawn exactly from π but their Markov dependence invalidates simple ranking arguments of the type used in Section 2.2 and we must seek more subtle solutions.

Besag and Clifford [**5**] describe two procedures that restore the ranking argument, despite the dependence in the chain. Both involve the ability to run a stationary Markov chain backwards, as well as forwards, in time. Thus, let $P(x, x')$ denote the (x, x') element of P and recall that, when a chain is stationary, its backwards transition probability matrix is Q, where Q has (x, x') element $Q(x, x') = \pi(x')P(x', x)/\pi(x)$. Often in MCMC, P is time–reversible with respect to π, in which case $Q = P$, but we do not insist on this property.

In Besag and Clifford's *parallel* procedure, positive integers r and n are chosen, dependent on the computational resources and the complexity of S: often we choose $r = 1000$ or 10000 and $n = 1000$. The chain is first run *backwards* from $x^{(1)}$ for r steps, according to Q, to obtain a state $x^{(0)}$, say. It is then run r steps forwards from this state, $n - 1$ times independently, to obtain states $x^{(2)}, \ldots, x^{(n)}$, say. Thus, if $x^{(1)}$ is a draw from π, then so are $x^{(0)}, x^{(2)}, \ldots, x^{(n)}$. More important, the procedure ensures that $x^{(1)}, \ldots, x^{(n)}$ are drawn *exchangeably* from π; that is, their joint distribution is permutation invariant. Exchangeability must then be inherited by the joint distribution of $u^{(1)}, \ldots, u^{(n)}$, for any particular test statistic u, and the rank of $u^{(1)}$ among $u^{(1)}, \ldots, u^{(n)}$ again has a uniform distribution, despite the dependence. Thus, the rank of $u^{(1)}$ again gives a valid p–value, as in Section 2.2. Note here that $x^{(0)}$ does not share the exchangeability and also that the theory would not apply if instead we ran backwards to a new $x^{(0)}$ on each of the $n - 1$ occasions.

In Besag and Clifford's *serial* procedure, integers r and n are again specified but also an integer v is chosen uniformly from $1, \ldots, n$. It is then arranged for $x^{(1)}$ to appear as the vth sample out of n in a single run of the chain, where samples are taken at gaps of r steps. This is achieved by running $r(v - 1)$ steps backwards and $r(n - v)$ steps forwards from $x^{(1)}$. Again, if $x^{(1)}$ is a draw from π and u is any particular test statistic, then the rank of the data value $u^{(1)}$ among all n values is uniformly distributed, after marginalizing over the distribution of v. The serial procedure is more untidy than the parallel version but generally has greater power for the same values of r and n because, on average, the recorded samples are at distance $\frac{1}{3}r(n + 1)$ from $x^{(1)}$, rather than at distance $2r$. Also, for a fixed run length rn, there is no inherent virtue in taking $r > 1$ in the serial test, except that subsampling tends to inhibit the occurrence of ties. Sequential versions of both the parallel and sequential procedures are described in Besag and Clifford [**6**] and these generally provide computational savings when there is no conflict between the model π and the observed test statistic $u^{(1)}$.

We must now negotiate the final complication in devising an MCMC procedure. Thus, although it is usually easy to produce an aperiodic P that respects stationarity with respect to π, irreducibility in a constrained state space S can be very difficult to achieve and, without this, π is not the *limiting* distribution of the chain. Strangely, as noted by Besag and Clifford [**5**] (Section 4), the validity of MCMC p–values does not in fact require irreducibility! If $S^{(1)}$ denotes the subset of S that can be reached from the data $x^{(1)}$ by repeated use of P, then the above arguments and the corresponding p–values still apply conditionally on the deficient state space $S^{(1)}$ and hence also marginally. Note that the reasoning depends critically on initializing the chain by $x^{(1)}$. Of course, if the loss of mobility is severe, then the power of the test against other models of interest may be seriously impaired; that is, it may be difficult or impossible to reject the postulated model for any $x^{(1)}$. Although one can devise toy examples in which the restriction to $S^{(1)}$ *increases* the power of the test against particular alternatives, this is unlikely to occur or at least have much impact in practice and one would generally prefer an effective irreducible P. Here, the term "effective" is added because it is trivial to achieve irreducibility in a finite state space S: one need merely include occasional Hastings proposals, chosen uniformly from a finite space that contains S, even though such moves may have negligible acceptance probabilities, in which case there is no actual improvement in mobility and the "remedy" is useless.

Note that Besag and Clifford [**5**] apply a reducible algorithm in testing whether an observed dataset $x^{(1)}$, whose elements $x^{(1)}_{ij} = \pm 1$ represent diseased and healthy plants on a two–dimensional array, is compatible with a finite–lattice Ising model having the same boundary values. The sufficient statistics for the two parameters in the model are $a = \sum x^{(1)}_{ij}$ and $b = \sum x^{(1)}_{ij} x^{(1)}_{i'j'}$, in which the summation for b is over all adjacent pairs of sites (i, j) and (i', j'). Hence the reference distribution π is uniform on the set of arrays S having the same values of a and b and the same boundary as $x^{(1)}$. A simple Metropolis algorithm with π as its stationary distribution is obtained by proposing swaps between the values at randomly selected pairs of interior sites and accepting those for which b, as well as a, is preserved. This algorithm, although reducible with respect to S, is sufficiently mobile in Besag and Clifford's application to reveal a conflict between the model and the data. Of course, the algorithm would not be valid in other MCMC contexts and it remains an open problem to devise an irreducible version.

In the next section, we return to Example 2 and to the construction of a corresponding irreducible chain but we emphasize here that reducible MCMC algorithms can still be useful for statistical tests. We make this point again in Section 4.

3 Testing for no three–way interaction in $I \times J \times K$ contingency tables

Recall that, in Example 2, we observe an $I \times J \times K$ contingency table $x^{(1)}$ and wish to assess its compatibility with the absence of three–way interaction, in which case the association between any two of the categorical variables is the same at all levels of the third, as in (1.2). Recall also that the corresponding reference distribution $\{\pi(x) : x \in S\}$ is given by (1.3), with S constrained to have all two–way margins x_{ij+}, x_{i+k} and x_{+jk} matching those of $x^{(1)}$. In order to obtain an MCMC exact p–value for (1.2), using any particular test statistic u, we require a transition probability matrix P for which π is a stationary distribution. We also prefer P to be irreducible with respect to S, as suggested in Section 2. Finally, it is convenient to restrict P to be time reversible with respect to π, so that P is also the backwards transition probability matrix of the chain.

We define an $I \times J \times K$ table to be a *move*, denoted by m, if its entries are all -1's, 0's and $+1$'s and its two–way margins are all equal to zero; we exclude $m \equiv 0$. Thus, if $x \in S$ and if m is a move, then $x + m$ is a table that preserves all two–way margins but may violate non–negativity. The simplest type of move m has all its elements $m_{ijk} = 0$, apart from $m_{i'j'k'} = m_{i'j''k''} = m_{i''j'k''} = m_{i''j''k'} = +1$ and $m_{i'j'k''} = m_{i'j''k'} = m_{i''j'k'} = m_{i''j''k''} = -1$, for two particular layers $i' \neq i''$, two particular rows $j' \neq j''$, and two particular columns $k' \neq k''$. We refer to such a move as a *basic move* and represent it diagrammatically by

$$\begin{array}{cc} + & - \\ - & + \end{array} \qquad\qquad \begin{array}{cc} - & + \\ + & - \end{array} \tag{3.1}$$

We write M for any particular finite set of moves, with the proviso that $m \in M$ implies that $-m \in M$. For reasons that will become apparent later, we denote the set of basic moves by M_2: this set features prominently in the sequel.

With any given M, we associate a transition probability matrix $P = P(M)$, defined operationally as follows. Let $x \in S$ denote the current state of the chain. Now choose m uniformly at random from M and define $x' = x + m$. If $x' \in S$, select x' as the next state of the chain with probability $\min\{1, \pi(x')/\pi(x)\}$; otherwise retain x. Clearly, P maintains π, is time reversible and corresponds to a single Metropolis step but P is not necessarily irreducible; indeed, it is possible that the chain is unable to leave its initial state $x^{(1)}$.

It is convenient to write M_* in place of M if its elements are irreducible with respect to S, by which we mean that any x and $x' \in S$ are connected via a path through S based on moves $m \in M_*$. Clearly, if we replace M by M_* in the definition of P, then P is irreducible with respect to S. We do not know of a parsimonious set M_* for arbitrary three–dimensional tables but below we discuss in detail the case $I = 2$. This is of practical importance in logistic regression and also leads to wider speculation.

3.1 The Rasch model and a result for $I = 2$. By focusing on $I = 2$, we are able to use known results for the celebrated Rasch [**16**] model: this has an enormous literature in educational testing. The practical setting is that each of J candidates attempts K items and scores 1 (correct) or 0 (incorrect) on each. Thus, the observed data form a $J \times K$ table of binary variables. The Rasch model postulates that all responses are independent and that the log–odds of a correct to an incorrect result in cell (j, k) can be written as the sum of a row j effect and a column k effect. It is easily shown that, in deriving a conditional test for the model, the corresponding reference distribution is uniform on the set of binary tables whose row and column totals tally with those observed. In practice, tables involving hundreds or thousands of candidates are quite common and the restriction to binary entries implies that the reference distribution is extremely difficult to deal with directly. Indeed, the Rasch model provided the original context for MCMC p–values in Besag [**3**]; see Besag and Clifford [**5**]. A key result, dating back at least to Ryser [**17**], is that any two binary tables with the same row and column totals can be connected by a sequence of moves of the type depicted in a single layer of (3.1).

The relevance of the Rasch model to the present paper is seen by recasting the above binary table as a $2 \times J \times K$ table in which the first layer is the original $J \times K$ table itself and the second is its complement. If we now imagine testing for no three–way interaction in the three–way table, then we are led to a task identical to the one above. In particular, the restriction to binary entries is satisfied automatically, since the layer totals are all unity. The equivalence in the tasks and the result at the end of the previous paragraph together imply that basic moves M_2 are irreducible in testing for the absence of three–way interaction in a $2 \times J \times K$ contingency table whose layer totals are all unity. In itself, this result is uninteresting but it also suggests that M_2 is irreducible for any $2 \times J \times K$ contingency table whose layer totals x_{+jk} are all positive. This becomes *Proposition 2* in Section 3.2 but first we establish by a simple example, due to John Skilling, that positivity of the x_{+jk}'s is not redundant. Thus, any basic move in the $2\times3\times3$ table

$$\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{array} \qquad\qquad \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array} \tag{3.2}$$

produces at least one negative element but the corresponding table with its layers interchanged has the same two–dimensional margins. We defer the case of arbitrary margins until Section 3.3.

3.2 Circuits and irreducibility for $I = 2$. Let $x^{(1)}$ denote a $2 \times J \times K$ contingency table with conditional space S. We follow Holst [**13**] and Sturmfels [**18**] (Ch. 13) in constructing a particular M_* that is irreducible with respect to S and has a one–to–one correspondence with the set of circuits in a bipartite graph on J and K nodes.

Thus, consider an undirected graph, with $J+K$ nodes, partitioned into a "row" set, indexed by $j = 1, \ldots, J$, and a "column" set, indexed by $k = 1, \ldots, K$. The graph is called *bipartite* on J and K nodes if each edge has a row node at one end and a column node at the other. A *circuit* of length $2v$ is then a sequence (j_1, k_1), (k_1, j_2), (j_2, k_2), $\ldots$, (j_v, k_v), (k_v, j_1), in which each element corresponds to a distinct edge of the graph. Any such circuit can be identified with a particular move, whose $2\times v\times v$ support is depicted in (3.3) up to permutations of rows and columns:

$$\begin{array}{c} \\ j_1 \\ j_2 \\ j_3 \\ \vdots \\ j_v \end{array} \quad \begin{array}{ccccc} k_1 & k_2 & k_3 & \ldots & k_v \\ + & 0 & 0 & \cdots & - \\ - & + & 0 & \cdots & 0 \\ 0 & - & + & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & + \end{array} \qquad \begin{array}{ccccc} k_1 & k_2 & k_3 & \ldots & k_v \\ - & 0 & 0 & \cdots & + \\ + & - & 0 & \cdots & 0 \\ 0 & + & - & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & - \end{array} \tag{3.3}$$

We now define M_v to be the set of all moves with support corresponding to these permutations; note that M_2 is the set of all basic moves, as before. Then we may choose (Diaconis and Sturmfels [**10**]),

$$M_* = M_2 \cup M_3 \cup \ldots \cup M_{\min(J,K)}\,. \tag{3.4}$$

Henceforth, M_* refers only to this specific set of moves. In the Appendix, we prove the following results, the second of which establishes the claim in Section 3.1.

Proposition 1 Any $m \in M_*$ can be decomposed into basic moves.

Proposition 2 If $x^{(1)}_{+jk} \geq 1$ for all j and k, then M_2 is irreducible for S.

3.3 Basic moves and irreducibility for arbitrary margins. We now allow the observed table $x^{(1)}$ to have some layer margins that are zero. Then, M_* in (3.4) is irreducible for S but M_2 may fail, as in (3.2). Nevertheless, we show that M_2 can still be used to construct an irreducible Markov chain on S. This is useful because M_* can be very awkward to work with directly, even for quite moderate values of J and K.

Suppose that $x \in S$ has the properties, (i) $x + m \notin S$ for all $m \in M_2$ and (ii) for some $v > 2$, there exists $m' \in M_v$ such that $x + m' \in S$. Let S' be the space of $2 \times J \times K$ tables with the same two–way margins as $x^{(1)}$ and having all non–negative entries except for at most a single -1. Then by *Proposition 1*, m' can be written as a sum of $v-1$ basic moves. In the Appendix, we prove the following

Proposition 3 If $x, x+m \in S$, where $m \in M_*$, then there exists a path, using moves in M_2, that connects x to $x+m$ and that does not leave S'.

Thus, there exists an ordering of the moves such that the corresponding path from x to $x+m$ remains in S' throughout.

Hence, by exploiting paths through S', we can construct an MCMC algorithm that is irreducible with respect to its state space S but whose transition probability matrix P' is built up only from moves in M_2 and does not require explicit identification of M_*. Note that the restriction to S' is important: a larger set might lead to paths that get "lost" in $S' \backslash S$. We now provide an operational definition of P' based on the above remarks.

Let $x \in S$ denote the current state of the chain and $x' \in S$ the subsequent state. Given x, first choose m uniformly at random from M_2 and set $x^* \leftarrow x+m$. If $x^* \notin S'$ then $x' \leftarrow x$. If $x^* \in S$, then x^* becomes a proposal and $x' \leftarrow x^*$ with probability $\min\{1, \pi(x^*)/\pi(x)\}$, else $x' \leftarrow x$. If $x^* \in S' \backslash S$, then draw another m at random from M_2 and update $x^* \leftarrow x^* + m$ if the new $x^* \in S'$, else do nothing. Repeat this last step until $x^* \in S$, in which case x^* is a proposal for x' and is treated as above.

It is easily established that P' corresponds to a single Metropolis step for the target distribution π. Note, in passing, that the algorithm is also applicable to tables with certain fixed or *structural* entries. Thus, if $x^{(1)}_{ijk}$ is a such an entry, one merely regards $x^{(1)}_{1jk}$ and $x^{(1)}_{2jk}$ as zeros during the course of the algorithm and restores their values in the eventual output.

4 Discussion

Our work was stimulated in part by preliminary versions of Diaconis and Sturmfels [**10**]. In the present context but for more general $I \times J \times K$ contingency tables, these authors demonstrate a one–to–one correspondence between a set of moves M that is irreducible for S and the Gröbner Basis (GB) of a certain polynomial ideal $\mathcal{I}_S$, where M now denotes a finite set of $I \times J \times K$ tables with arbitrary integer entries and zero two–dimensional margins. By identifying each position (i,j,k) in the table with an indeterminate z_{ijk}, one obtains an equivalence between a table x and a monomial in indeterminates $z_{111}, \ldots, z_{IJK}$:

$$x \longleftrightarrow z_{111}^{x_{111}} \, z_{112}^{x_{112}} \, \ldots \, z_{IJK}^{x_{IJK}} \,. \tag{4.1}$$

Furthermore, any ordering $\prec$ on $\mathcal{N}^{I \times J \times K}$ produces an ordering on monomials. Then, if z^x denotes the monomial in (4.1), one defines $z^x \prec z^{x'}$ if $x \prec x'$. For example, in the case of lexicographic ordering, $x \prec_{lex} x'$ if the leftmost nonzero entry in $x - x'$ is positive. An ordering on monomials induces an ordering on the indeterminates.

Consider a monomial ordering and define

$$m^+ = \max(m, 0)\,, \qquad m^- = \max(-m, 0)\,,$$

where the maxima are taken elementwise. Then, following Sturmfels [**18**] (Ch. 5), a finite set M is irreducible for S if and only if $\{x^{m^+} - x^{m^-} : m \in M\}$ generates $\mathcal{I}_S$ with respect to that ordering, where $\mathcal{I}_S$ denotes the ideal generated by $\{z^x - z^{x'} : x, x' \in S\}$. Although, in principle, this identifies an irreducible M, explicit computation is usually prohibitive. Even the case $I = J = K = 3$ is quite taxing (Diaconis and Sturmfels [**10**]).

For $I = 2$, Sturmfels [**18**](Ch. 13) shows that $\{x^{m^+} - x^{m^-} : m \in M_*\}$ is a GB for all monomial orderings simultaneously, where M_* is defined as in (3.4). As a consequence, M_* is irreducible for S, irrespective of the values of the two–dimensional margins $x^{(1)}_{ij+}$, $x^{(1)}_{i+k}$ and $x^{(1)}_{+jk}$ of an observed table $x^{(1)}$.

There is also an interesting link between our results and those in Jacobson and Matthews [**14**], which describes the use of MCMC in generating random $K \times K$ Latin squares. Thus, let x denote a $K \times K \times K$ table of 0's and 1's having *all* its two–dimensional margins equal to unity. Then there is an equivalent Latin square in which symbol k appears in those positions (i, j) for which $x_{ijk} = 1$. Jacobson and Matthews [**14**] show, by methods that differ from ours, that all $K \times K$ Latin squares, when rewritten in their three–dimensional form, are also connected by basic moves, provided that again one allows intermediate visits to tables with a single -1 entry. It is therefore natural to speculate that *Proposition 3* extends to any $I \times J \times K$ contingency table, subject to fixed positive two–dimensional margins. Note here that examples can be constructed to demonstrate that some form of positivity condition is required. Although we have not proved our conjecture, we recall that irreducibility is not obligatory for the validity of MCMC p–values, so that we are at least assured that the use of basic moves provides a legitimate test for no three–way interaction in $I \times J \times K$ tables.

Finally, in assessing almost any model of interest for higher–dimensional contingency tables, the corresponding state space S constrains particular marginal totals to match those in the observed data. Then, again if one is willing to forsake irreducibility, it is easy to extend the notion of basic moves to construct an MCMC exact test for the model. As an example, suppose that $x^{(1)}$ denotes an observed $I \times J \times K \times L$ table with entries $x^{(1)}_{ijkl}$ and that we wish to test for the "four–cycle" in which cell (i, j, k, l) has multinomial probability (cf. (1.2))

$$p_{ijkl} \;=\; \theta_{ij}\, \phi_{jk}\, \psi_{kl}\, \xi_{li}\,.$$

Then sufficient statistics for the unknown parameters are the corresponding two–dimensional margins, $x^{(1)}_{ij++}$, $x^{(1)}_{+jk+}$, $x^{(1)}_{++kl}$ and $x^{(1)}_{i++l}$, so that we require a set of moves M for which these are preserved. If we now allow moves based on four–dimensional hypercubes of ± 1s, we obtain 30 solutions of which eight are linearly independent, compared with the single previous one in (3.1). We have used similarly–constructed moves for analyzing various forms of interaction in up to six–dimensional contingency tables, with parametrizations that are not necessarily symmetric as in the examples here. Despite the presumed absence of irreducibility and hence restricted mobility, the MCMC algorithms have been successful in the sense that they have rejected hypothesized models with annoying regularity! Also the MCMC computations have confirmed the frequent irrelevance of standard asymptotic approximations in practice. Despite these comments, a practicable method of identifying irreducible sets of moves in constrained state spaces would be a considerable advance.

Acknowledgements

We acknowledge helpful discussions with Simon Byers, Persi Diaconis, Mark Jacobson and Thomas Richardson. The first author received financial support from the National Research Center for Statistics and the Environment at the University of Washington.

Appendix

Proof of Proposition 1

We need to prove that any $m \in M_v$, where $2 \leq v \leq \min(J,K)$, can be regarded as a sum of moves from M_2. It is sufficient to consider a move m having support as in (3.3) and for this we can write an explicit decomposition, $m = m^1 + \ldots + m^{v-1}$, where $m^s \in M_2$ corresponds to the circuit (j_1, k_s), (k_s, j_{s+1}), (j_{s+1}, k_{s+1}), (k_{s+1}, j_1). Hence, the result is verified.

Proof of Proposition 2

Suppose that $x_{+jk} \geq 1$ for all j and k and that $x, x+m \in S$, where $m \in M_*$. We need to show that there exists a sequence of moves in M_2 from x to $x+m$ that maintains non–negativity in all intermediate tables.

Define another $2 \times J \times K$ table y as follows. If $m_{ijk} = 1$, set $y_{ijk} = 0$; if $m_{ijk} = -1$, set $y_{ijk} = 1$; if $m_{ijk} = 0$ and $x_{1jk}x_{2jk} = 0$, set $y_{ijk} = \min(1, x_{ijk})$; otherwise, set $y_{1jk} = 1$ and $y_{2jk} = 0$. Then y has binary entries and $y_{+jk} = 1$ for all j and k. Thus, y corresponds to a Rasch table and so does $y+m$, with the same two–dimensional layer, row and column totals as y: non–negativity is preserved because $y_{ijk} = 1$ whenever $m_{ijk} = -1$. It follows that $y+m$ can be reached by a sequence of moves in M_2, each of which maintains non–negativity. Moreover, $x_{ijk} \geq y_{ijk}$, so that the same sequence of moves applied to x produces $x+m$ and again preserves non–negativity.

Proof of Proposition 3

It is sufficient to consider moves m having support as in (3.3), for which we have the decomposition used in the proof of *Proposition 1*. Since $x+m \in S$, the only entry that can be negative when applying the M_2–move that corresponds to the circuit (j_1, k_s), (k_s, j_{s+1}), (j_{s+1}, k_{s+1}), (k_{s+1}, j_1) is in position $(1, j_1, k_{s+1})$, as row j_{s+1} and column k_s are not modified subsequently. Furthermore, the $(1, j_1, k_{s+1})$ element can at worst be -1, since it is modified only once subsequently.

References

[1] Barnard, G. A. *Discussion on paper by M.S. Bartlett*, J. R. Statist. Soc. B **25** (1963), 294.
[2] Bartlett, M.S. *Contingency table interactions*, Suppl. J. R. Statist. Soc. **2** (1935), 248–252.
[3] Besag, J. *Unpublished workshop presentation* (1983), 44th Session of the Int. Statist. Inst., Madrid.
[4] Besag, J. *Simple Monte Carlo p–values*, Proc. of Interface 90, Springer–Verlag, New York, 1992, pp. 158–162.
[5] Besag, J., Clifford, P. *Generalized Monte Carlo significance tests*, Biometrika **76** (1989), 633–642.
[6] Besag, J., Clifford, P. *Sequential Monte Carlo p-values*, Biometrika **78** (1991), 301–304.
[7] Besag, J., Diggle, P.J. *Simple Monte Carlo tests for spatial pattern*, Appl. Statist., **26** (1977), 327-333.
[8] Cox, D., Little, J., O'Shea, D. *Ideals, Varieties and Algorithms*, Springer–Verlag, New York, 1996.
[9] Cox, D. R., Snell, E. J. *Analysis of Binary Data*, Chapman Hall, London, 1989.
[10] Diaconis, P., Sturmfels, B. *Algebraic algorithms for sampling from conditional distributions*, Ann. Statist. **26** (1998), 363–398.

[11] Fisher, R. A. *On the mathematical foundations of theoretical statistics*, Phil. Trans. R. Soc. London **A 222** (1922), 309–368.

[12] Guo, S. W., Thompson, E.A. *Monte Carlo estimation of mixed models for large complex pedigrees*, Biometrics **50** (1994), 417–432.

[13] Holst, C. *Item response theory*, Technical Report (1995), Danish Educational Research Institute, Copenhagen.

[14] Jacobson, M. T., Matthews, P. *Generating uniformly distributed random latin squares*, J. Combinat. Des. **4** (1996), 405–437.

[15] Patefield, W. M. *Algorithm AS 159. An efficient method of generating random* $r \times c$ *tables with given row and column sums*, Appl. Statist. **30** (1981), 91–97.

[16] Rasch, G. *Probabilistic Models for some Intelligence and Attainment Tests*, Danish Educational Research Institute, Copenhagen, 1960.

[17] Ryser, H. J. *Combinatorial Mathematics*, vol. 14 of The Caurus Mathematical Monographs. The Mathematical Association of America, 1963.

[18] Sturmfels, B. *Gröbner Bases and Convex Polytopes*, Amer. Math. Soc., Providence, RI, 1996.