

RECONSTRUCTION OF CONTINGENCY TABLES WITH MISSING DATA

CLAUDIA TEBALDI & MIKE WEST
ISDS, DUKE UNIVERSITY

REVISED VERSION: NOVEMBER 23, 1998

We describe and illustrate approaches to Bayesian inference in partially observed contingency tables. Key examples include problems of observation of only selected marginal totals, when interest lies in inference about unobserved cell counts and underlying cell probabilities. A main focus of this work is the presentation and illustration of a novel and efficient simulation algorithm for imputation of missing cell counts and its incorporation into Bayesian analyses of partially observed tables. We present and illustrate the algorithm in a context of two-way tables, where we also introduce a new and flexible class of prior distributions for parameters in saturated log-linear models. Illustration is provided using data arising from a NISS transportation policy research project, and a second data set used to compare the approach with existing and standard algorithms. This second example illustrates the practical efficacy of the new imputation, and the major extent to which it dominates existing methods from an applied point of view.

Keywords: *Bayesian inference; Imputation; Missing data; Log-linear models; Markov chain Monte Carlo.*

Claudia Tebaldi, PhD, is research fellow at the National Center for Atmospheric Research, Boulder CO. Mike West is Professor and Director of the Institute of Statistics and Decision Sciences, Duke University. The research reported here was partially supported by NSF grant DMS-9313013 to the National Institute of Statistical Sciences. The corresponding author is Claudia Tebaldi at NCAR, Climate Analysis Section/Geophysical Statistics Project, Boulder, CO, 80307.

1. INTRODUCTION

Consider a two-way contingency table with r rows and c columns, and having uncertain cell probabilities $\pi = \{\pi_{ij} : i = 1, \dots, r; j = 1, \dots, c\}$. Suppose the canonical multinomial sampling context; that is, we expect to observe cell counts $x = \{x_{ij} : i = 1, \dots, r; j = 1, \dots, c\}$ out of a total of n cases under a multinomial sampling model, and are interested in inferences on π . Suppose now that the full cell counts are unrecorded: only the row and column totals are recorded. Our problem is now to derive posterior inferences on π in this missing data context. Related problems begin with an inferential interest on the unobserved cell counts themselves: that is, infer the cell counts conditional on observed values of the row and column totals and in the context of uncertainty about cell probabilities. More broadly, we are interested in problems of inference about π and missing cell counts in contexts of partial observation of the contingency table recorded in terms of sums of selected cell counts; the margins-only case is a key example. Our focus and examples here concern two-way tables, although the applied imputation problems, and the solutions proposed, are relevant to multi-way tables in general.

Problems with this kind of structure arise in diverse areas of application such as in the computation of significance levels (Agresti, 1992; Diaconis and Efron, 1985; Smith and al., 1996) and, of more interest here and much more practically relevant, in socio-economic and demographic studies that involve and rely on survey and census data representing differing levels of aggregation (hence marginalisation) of population characteristics. Much of the activity in these latter areas has been referred to as micro-simulation (Fienberg, 1997), and it is the development of micro-simulation methods in transportation policy research areas (at NISS) that partially motivated the current work.

Bayesian inference in this context is in principle straightforward: we simply compute posterior distributions for the unobserved cell counts and cell probabilities jointly. In practice this may be addressed using the standard tools of Markov chain Monte Carlo (MCMC) simulation, and this requires creativity in dealing with simulations of the missing cell counts from appropriate conditional posterior distributions. The main contribution of this article is the introduction of a novel simulation algorithm for this component of the problem. The algorithm is a version of more general algorithms introduced by the authors in a quite different context, that of origin-destination flow estimation in problems of network inference (Tebaldi and West 1998). Other versions of this algorithm apply in contingency table problems with essentially arbitrary patterns of missing and aggregated data, and so are of more general relevance to the micro-simulation area. The specific context of tables with observed margins discussed here is important in practice and also a useful special vehicle for presentation. In addition to this primary focus on algorithms for imputation of missing data, our development also involves data analysis using a flexible class of priors for parameters in

saturated log-linear models of two-way table cell counts. Our examples demonstrate implementation and efficacy of the approach, and that it dominates existing, standard approaches to missing cell count imputation from an applied perspective.

2. ANALYSIS FRAMEWORK

We begin with some notation. The full set of counts x is assumed to be represented as an $r.c$ -column vector created by concatenating the rows of the table, i.e.,

$$x = (x_{1,1}, x_{1,2}, \dots, x_{1,c}; \dots; x_{r,1}, x_{r,2}, \dots, x_{r,c})'$$

Denote the row and column totals by x_{i+} and x_{+j} for $i = 1, \dots, r$ and $j = 1, \dots, c$, respectively, and

$$y = (x_{1+}, \dots, x_{r+}; x_{+1}, \dots, x_{+,c-1})'$$

for the $(r + c - 1)$ -column vector of row and column totals with that of the final column deleted. Notice that we can write

$$y = Ax$$

where A is an $(r + c - 1) \times (rc)$ -matrix of zeros and ones that simply selects rows and columns for summation, i.e., A has the block form

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_r \\ J & J & \cdots & J \end{pmatrix}$$

where each A_i is an $r.c$ -matrix of zeroes with elements 1 in the i^{th} row and 0 elsewhere, and J is a $(c - 1).c$ -matrix with elements $J_{ii} = 1$ for $i = 1, \dots, c - 1$ and 0 elsewhere.

In the margins-only context described in the introduction, we have a problem in which, instead of observing the full data x , we observe the partial/aggregate data y . Notice that, from a purely algebraic viewpoint, we have an under-determined system of linear equations $y = Ax$ in which x is higher dimensional than y . Note also that, by construction, the matrix A is of full rank $r + c - 1$.

We work under the standard assumption that the multinomial sampling model is derived from the collection of conditionally independent Poisson models $Po(\cdot | \lambda_{ij})$ for cell counts x_{ij} , and so the cell probabilities are $\pi_{ij} = \lambda_{ij} / \tau$ with $\tau = \sum_{i,j} \lambda_{ij}$. Under a log-linear model, as used below, inference on λ under the product Poisson model is equivalent to inference on π in the multinomial model; we work in terms of the former for tractability. Assume a specified prior $p(\lambda)$ for the set of cell rates $\lambda = \{\lambda_{ij} : i = 1, \dots, r; j = 1, \dots, c\}$. Each applied context requires its own prior, or class of priors, of course; our development below introduces and discusses a flexible prior modelling framework. For now, we simply assume a specific prior $p(\lambda)$ is specified.

Often an application will involve additional, or prior, observations from the full joint distribution that the contingency represents, so we include that possibility by extending the framework, in a minor way. That is, in addition to the observed marginal data y , we have additional sample data available from the full contingency table, an additional set of observations $z = \{z_{ij} : i = 1, \dots, r; j = 1, \dots, c\}$, with a total m , drawn from the multinomial on cell probabilities π and independently of x . This essentially generalises the margins-only framework, which is recovered as the special case in which $m = 0$. In some applications, such data is available in which case it may be used to update relatively uninformed, even improper priors for cell probabilities to proper posteriors in advance of analysing the margins-only data y . We refer to this additional data as the collateral table as it provides additional, collateral information on the underlying multinomial distribution.

In Section 4 below we introduce specific classes of priors $p(\lambda)$ and develop the resulting posterior analysis using MCMC (Markov chain Monte Carlo) iterative simulation methods. The analysis iteratively re-simulates values of the model parameters and the missing data elements in x , so that an algorithm to resample sets of x values is critical. We now present and describe a novel algorithm for this.

3. IMPUTING CELL COUNTS

Consider the context of known model parameters λ , observed margins y and additional complete data z . Imputation of a new set of the missing cells counts x involves sampling from the conditional posterior distribution

$$p(x|z, y, \lambda)$$

or the implied sets of conditional distributions for selected subsets of elements of x . Evidently, $p(x|z, y, \lambda) = p(x|y, \lambda)$ and this is supplied by imposing the set of linear constraints $y = Ax$ on the underlying Poisson sampling model $p(x|\lambda)$. The resulting singular distribution has a complicated form, and simulation of new values for x must generally be done in a Gibbs sampling format, one value at a time conditional on the remaining values. Developing this is the focus of this section.

The approach is a version of the origin-destination flow framework developed in Tebaldi and West (1998) for a quite different problem. The discussion below translates that context to the current framework and describes the resulting algorithm. We begin by exploiting the algebraic structure of the problem and then deducing the sequence of univariate conditional posterior distributions for elements of x .

Recall the algebraic structure of the margins only data $y = Ax$ with A defined in Section 1. As A is full rank the columns may be reordered so that the resulting matrix has the form $A = [A_1, A_2]$ where A_1 is a non-singular $(r + c - 1) \cdot (r + c - 1)$ matrix. This reordering is

efficiently and automatically done using elements of the QR decomposition of the original A computed using the Householder successive reflection procedure (as implemented, for example, in the `qr()` routine in S-Plus); see Theorem 1 and ensuing discussion of Tebaldi and West (1998). Suppose from here on that A is reorganised this way. Then, similarly reordering the elements of the x vector and conformably partitioning as $x' = (x'_1, x'_2)$, it follows that $x_1 = A_1^{-1}(y - A_2x_2)$. This simply reflects the fact that, in the presence of knowledge of the margins y , there are really only $(r-1)(c-1)$ free cell elements as represented by the $(r-1)(c-1)$ -vector x_2 . Given these in addition to y implies the $(r+c-1)$ -vector x_1 is known. Hence our attention is focused on sampling algorithms for the x_2 vector alone, whose value implies x_1 and hence a draw for the full vector x . We sample elements of x_2 individually as now detailed.

It is immediate that the full conditional posterior density for the vector x_2 has the form

$$p(x_2|z, y, \theta) \propto \prod_{i,j} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}$$

and with support defined by $x_{ij} \geq 0$ for all i, j . Hence, for any elements x_{ij} of x_2 we have

$$p(x_{ij}|z, y, \theta, x_2 \setminus x_{ij}) \propto \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!} \prod_{h,k} \frac{\lambda_{hk}^{x_{hk}}}{x_{hk}!}$$

where the product is taken over elements of x_1 only with values replaced by $x_1 = A_1^{-1}(y - A_2x_2)$, and where $x_2 \setminus x_{ij}$ is the vector x_2 with x_{ij} removed. The support of this univariate conditional posterior is identified through the linear constraints on x_{ij} , namely by $x_{ij} > 0$ and $x_{hk} \geq 0$ for all elements in the product. By directly evaluating these constraints and identifying their intersection we can deduce the support of the above conditional distribution and hence identify the unnormalised conditional posterior. This is discussed in Tebaldi and West (1998, Sections 2.3,2.6) where it is also shown (Appendix) that the support is a bounded and connected subset of non-negative integers, a result that is important in a direct and computationally efficient simulation analysis: we sequence through the elements x_{ij} of x_2 in turn, sampling new values according to the following Metropolis-Hastings scheme.

At each iteration, a candidate value of x_{ij} is generated from a specified proposal distribution, and accepted or rejected according to the usual Metropolis-Hastings acceptance probability (Tierney 1994). Specifically, assume we have specified a fixed proposal distribution with density $q_{ij}(\cdot)$ for element x_{ij} . A candidate value x_{ij}^* is drawn from this proposal, and accepted with probability

$$\min \left[1, \frac{p_{ij}(x_{ij}^*)q_{ij}(x_{ij})}{p_{ij}(x_{ij})q_{ij}(x_{ij}^*)} \right]$$

where x_{ij} is the current, most recently sampled value, and $p_{ij}(\cdot)$ is the unnormalised conditional posterior above, i.e., proportional to $p(x_{ij}|z, y, \theta, x_2 \setminus x_{ij})$. Note that this will be

evaluated only at candidate draws, so that, for a given proposal distribution, it is not necessary to either identify the actual support of the true conditional posterior nor to evaluate it completely across the support, so leading to efficiencies in sampling. If we use a proposal distribution whose support includes values of x_{ij} lying outside the posterior support, candidate draws in the illegal region will simply be rejected. Obvious choices of proposal distributions include uniforms and Poissons constrained to an identified finite range. From the algebraic structure of the network equations $y = Ax$ it is possible to identify gross bounds on each x_{ij} in advance of analysis, so that a suitable range for the proposal distribution can be computed. Details follow those of Tebaldi and West (1998, Section 3.1), and an automatic algorithm is easily coded to identify the gross bounds. Some technical elaboration is given in the appendix here. Our preferred choice of proposal distribution is the Poisson truncated to this identified range, i.e., taking $q_{ij}(x_{ij})$ to be $Po(x_{ij}|\lambda_{ij})$ truncated to the identified range, where λ_{ij} is evaluated at the current values of the model parameters θ . Experiments with both this and the alternative uniform proposals have been satisfactory in terms of reasonable acceptance rates and convergence, with the two being essentially undistinguished in most cases. Illustration in the next section is based on the Poisson proposal version. Note that both generation of candidate values and evaluation of the acceptance probabilities are essentially trivial.

As mentioned above, this development is very closely related to that of Tebaldi and West (1998) in a quite different framework, and readers interested in technical details will benefit from exploration of technical appendices in that article. In particular, we note that the Markov chain convergence theory of that paper applies here without modification, and this assures us that the Gibbs/Metropolis sampling algorithm described here does in fact converge to a stationary distribution that is the required joint posterior distribution.

4. PRIOR SPECIFICATION AND FULL POSTERIOR SAMPLING

The following analysis assumes a class of priors for λ induced by priors for the structural parameters in the representation

$$\lambda_{ij} = \mu\alpha_i\beta_j\gamma_{ij}$$

subject to the aliasing constraints $\alpha_1 = \beta_1 = 1$, $\gamma_{i,1} = 1$ for each $i = 1, \dots, r$ and $\gamma_{1,j} = 1$ for each $j = 1, \dots, c$. This is evidently equivalent to a standard, saturated log-linear model with marginal effects parameters $\log \alpha_i$ and $\log \beta_j$, and with interactions $\log \gamma_{ij}$. Write $\alpha = \{\alpha_i : i = 2, \dots, r\}$ for the row margin parameters, $\beta = \{\beta_j : j = 2, \dots, c\}$ for the column margin parameters, and $\gamma = \{\gamma_{ij} : i = 2, \dots, r; j = 2, \dots, c\}$ for the interaction parameters. Further, write $\theta = \{\mu, \alpha, \beta, \gamma\}$ for all parameters. A relatively unstructured class of priors for the λ_{ij} is induced by exchangeable/independence assumptions implicit in

the joint prior

$$p(\theta) = p(\mu, \alpha, \beta, \gamma) = p_m(\mu) \left\{ \prod_i p_a(\alpha_i) \right\} \left\{ \prod_j p_b(\beta_j) \right\} \left\{ \prod_{i,j} p_g(\gamma_{i,j}) \right\}$$

for some specified marginal prior densities $p_m(\cdot)$, $p_a(\cdot)$, $p_b(\cdot)$ and $p_g(\cdot)$. Often these may be taken as conditionally conjugate gamma densities or, as we assume additional sample data z is available from the full contingency table, the priors may be taken as reference improper priors. Similar priors were exploited in a related context in West (1994). We therefore have a full joint distribution for the two data sets given by

$$p(x, z | \theta) = p(x | \lambda) p(z | \lambda) = \prod_{i,j} Po(x_{ij} | \lambda_{ij}) Po(z_{ij} | \lambda_{ij}),$$

the product being over all row and column indices. Were x fully observed, this would provide the likelihood function for θ . This distribution has key conditionals, as follows.

- The conditional for x is given by

$$p(x | z, y, \theta) = p(x | y, \theta) \propto \prod_{i,j} Po(x_{ij} | \lambda_{ij})$$

subject to the linear constraints $y = Ax$, as described in the previous section.

- The conditional for θ is given by

$$p(\theta | z, y, x) \equiv p(\theta | z, x) \propto p(\pi) \prod_{i,j} Po(x_{ij} | \lambda_{ij}) Po(z_{ij} | \lambda_{ij}),$$

simply observing that knowledge of the full tables of counts x and z reduces inference on θ to that of standard prior:posterior updating in the standard product Poisson model.

Iterative MCMC simulation of the joint posterior may be implemented by sequencing through elements of the parameter θ and missing cell counts x in turn, at each step sampling the quantities in question from the appropriate conditional posterior given the latest sampled values of all required conditioning variates. In the previous section we have discussed the approach to sampling elements of x one at a time. The complementary sampling exercises involve drawing new values of the model parameters θ . As in West (1994), the current structure makes this immediately accessible using Gibbs sampling, as follows.

For any set of elements ϕ of θ , write $\theta \setminus \phi$ for all parameters but ϕ . Then we note:

- $p(\mu | z, x, \theta \setminus \mu) \propto p_m(\mu) \mu^M e^{-\mu m}$ where $m = 2 \sum_{i,j} \alpha_i \beta_j \gamma_{ij}$ and $M = x_{++} + z_{++}$. Hence a gamma prior for μ induces a gamma conditional posterior, as does the standard improper reference prior $p_m(\mu) \propto \mu^{-1}$.

- For each $i = 2, \dots, r$, $p(\alpha_i | z, x, \theta \setminus \alpha_i) \propto p_a(\alpha_i) \alpha_i^{A_i} e^{-\alpha_i a_i}$ where $a_i = 2\mu \sum_j \beta_j \gamma_{ij}$ and $A_i = x_{i+} + z_{i+}$. Further, the α_i are conditionally independent over rows i . Hence a gamma form for $p_a(\cdot)$ induces gamma conditional posteriors for each of the α_i , as does the standard improper reference prior $p_a(\alpha_i) \propto \alpha_i^{-1}$.
- For each $j = 2, \dots, c$, $p(\beta_j | z, x, \theta \setminus \beta_j) \propto p_b(\beta_j) \beta_j^{B_j} e^{-\beta_j b_j}$ where $b_j = 2\mu \sum_i \alpha_i \gamma_{ij}$ and $B_j = x_{+j} + z_{+j}$. Further, the β_j are conditionally independent over columns j . Hence a gamma form for $p_b(\cdot)$ induces gamma conditional posteriors for each of the β_j , as does the standard improper reference prior $p_b(\beta_j) \propto \beta_j^{-1}$.
- For each $i = 1, \dots, r$ and $j = 2, \dots, c$, $p(\gamma_{ij} | z, x, \theta \setminus \gamma_{ij}) \propto p_g(\gamma_{ij}) \gamma_{ij}^{G_{ij}} e^{-\gamma_{ij} g_{ij}}$ where $g_{ij} = 2\mu \alpha_i \beta_j$ and $G_{ij} = x_{ij} + z_{ij}$. Further, the γ_{ij} are conditionally independent over rows and columns. Hence a gamma form for $p_g(\cdot)$ induces gamma conditional posteriors for each of the γ_{ij} , as does the standard improper reference prior $p_g(\gamma_{ij}) \propto \gamma_{ij}^{-1}$.

In contexts with no additional data z , the z terms in the above are all zero and the 2 factors do not appear in m , a_i , b_j and g_{ij} .

Hence, Gibbs sampling analysis may be directly implemented for the model parameters by sequencing through the above set of conditional posteriors, resampling new values in the standard format. At each iteration, new values of the missing cell counts x are then sampled conditional on the model parameters, as detailed in the previous section.

5. ILLUSTRATION

An initial illustrative example concerns problems of reconstructing the joint distribution of selected socio-demographic variables at a fine geographical level, and is a simple illustration of the kinds of reconstruction/imputation issues arising in socio-economic micro-simulation studies. The particular context that motivated our algorithmic development concerns imputation of household demographics in studies of urban transportation. Data arising in a NISS study based on surveys and other records of areas of Napa County, California, include US Census data at the so-called block group level. A typical US Census block group covers an area of approximately 250 households/families and the data base of interest here is PUMA (Public Use Micro Area) level: it covers the PUMA 900 area of Napa County, a region of 109 block groups, comprising 28,621 households. The data base provides a full set of univariate marginal totals for a collection of socio-demographic variables at the block group level.

For example and illustration here, we focus on two key variables:

- family income, denoted by I and divided into 5 levels (lower bounds, in thousands of dollars, at 0, 17.5, 30.0, 42.5, 60.0), and

- number of workers in the family, denoted by W and divided into 4 levels (0, 1, 2, 3 and ≥ 4).

The actual sizes of the blocks in the data set vary from 8 to 993 households, the median size being 233, and they represent widely variable marginal totals for the selected variables I and W . Two block groups are selected for study here: Block 9, with 241 households, and Block 28, with 289 households. The marginal totals on variables I and W are shown in Figures 1 and 2 for these two block groups, respectively. We perform separate analyses of these two data sets separately, to illustrate the algorithm and various features. In order to study and assess the performance of the imputation algorithm, we adapt additional data from the survey as providing the collateral, fully observed contingency table z in the previous sections. The census data includes a representative summary of 5% of the entire PUMA area, which we rescale to provide a full table of approximately the same sample size (m) as each of the selected blocks. This is used as the data z in each of two separate analyses, one for each of Blocks 9 and 28, respectively. Figure 3 represents the cells of the 5×4 collateral table z , with different degrees of shading (darker for higher probability). Figure 4 shows the two sets of marginal totals for the table z , that may be compared with the observed margins of the two blocks under examination: Block 28 has marginal totals whose distributional shapes are consistent with those of the collateral table z , whereas those of Block 9 are quite different.

Experimentation and a range of empirical studies have indicated that the MCMC analysis typically converges reasonably rapidly; the results here are based on runs of 15000 iterations, of which the initial 500 are discarded as the burn-in stage. Such analyses are quite efficient computationally, running in a matter of seconds on current, cheap desk-top work-stations for small tables such as in this example. This is important in terms of potential extensions and applicability to more complex problems: larger tables, and multidimensional tables, the processing of which might be partitioned in a series of stages, each based on a simple two-way table. The need for a starting point (a feasible solution for the cell counts consistent with the given margins) is easily satisfied. After computing different starting points by heuristic inspection, an automatic initialisation procedure is easily constructed and built into the algorithm. Details of the initialisation are summarised in the appendix. This, and the MCMC methodology, are available in Fortran software on request from the first author.

The results briefly summarised here are given for analyses of the two Blocks 9 and 28 separately. In each analysis, the full collateral table z is used as discussed in the earlier methodology section. The analyses are monitored using standard graphical displays of iterates of selected parameters and cell counts to informally assess convergence, and results are summarised for display in terms of raw histograms of various margins of posterior

distributions. Posterior margins of interest are those for underlying cell probabilities π_{ij} in each case, as well as the posterior predictive distributions for the missing cell counts in each block table. Figures 5 and 6 display the former: posterior margins and summaries for the two underlying sets of multinomial probabilities. These differ markedly between the two block groups, consistent with the observed disparities between the marginal counts of the two blocks, and in spite of the common data provided by the collateral table z . In each case, the approximate posterior means of the π_{ij} are indicated above the histograms, and the observed counts from the z table appear on the lower axis for comparison. The resulting inferences on several of the π_{ij} suggest obviously different values than indicated by the table z alone, adapting to the additional margins-only data. Additional insights into the posteriors, and comparisons with the fully observed collateral table z , are gained from Figures 7 (for Block 9) and 8 (for Block 28). These display three sets of margins for the two variables I and W. The first row in these figures displays the observed margins the z table; the second row displays the actual observed counts in the two blocks; the third row presents the posterior predictive margins estimates, computed simply as the Monte Carlo averages of the margins of samples of tables imputed in the MCMC analysis. The difference between the two blocks is very evident here. Block 9 has observed margins that differ substantially from those of the collateral table z , and as a result produces posterior distributions that are significantly different from those in the case of Block 28, the latter having observed margins in much closer concordance with those of z , as displayed.

6. COMPARISON WITH STANDARD METHODS

Existing approaches to imputation of missing cell counts given observed margins or other linear constraints include variants of local-move iterative simulation methods (Diaconis and Sturmfels 1998; Fienberg 1997; Nobile 1997). The following is an example of such standard Metropolis-Hastings methods: at each iteration, generate a cell count configuration that perturbs the current cell counts by randomly selecting a local 2×2 sub-table, and changing it by adding 1 (or some other fixed integer) to the diagonal counts, while subtracting 1 from the off-diagonals. This maintains the fixed margins and produces a proposal for new cell counts that may be accepted or rejected in the usual way. Theory of this, as an example of a vastly more general mathematical framework, appears in Diaconis and Sturmfels (1998). Note that such local-move algorithms are easily incorporated into a Bayesian analysis; in our framework above, simply replace the imputation module with such an alternative. Nobile (1997), in as yet unpublished work, implemented such methods.

We use the data given in Table 2, page 368 of the paper by Diaconis and Sturmfels (1998) to illustrate the efficacy of our implementation and compare it to the standard Metropolis-Hastings method of local moves, referred to in the same paper.

Table 1: An example.

	Hair color				
Eye color	Black	Brunette	Red	Blonde	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

Table 1 represents the joint distribution of a sample of 592 women with respect to the color of their hair and eyes. We consider the margins of Table 1 as our y vector and the content of the cells as missing data, using the known x_{ij} values as starting points for simulation analysis. A comparison of the two imputation approaches – our new algorithm and the standard local-move algorithm – is carried out in the context of the earlier described model and prior structure; the analyses differ only through the simulation/imputation algorithm, or module, for the missing cell counts.

Iteration by iteration, the local-move algorithm is obviously faster than the new global Metropolis method, but suffers very significantly by comparison in terms of the very high dependencies between consecutive sample tables and corresponding parameter values. For example, consider the sample autocorrelation function of simulated values of selected elements of the λ parameter vector. Figure 9 presents the sample autocorrelation functions for one randomly chosen λ parameter in a long run of the two MCMC analyses. In fact, we saved only every tenth sample value, so that the Lag index in the autocorrelations displayed actually represents ten steps. Hence the extremely high and persistent dependencies in the upper frame imply even higher dependencies per iteration. This graph clearly illustrates the very high correlations inherent in output streams from local-move methods, and that the new algorithm is, in stark contrast, far better conditioned.

Figure 10 displays the posterior margins for the set of cell counts as they appear after 10,000 iterations of the new global algorithm, which, by standard diagnostic tools (e.g., Best et al 1995) has already reached convergence. Figure 11 shows the same margins after the same number of iterations as they result from running the local-move algorithm, where the apparent multimodalities arise through transient behaviour as this chain is far from converged. This example simply illustrates the practical dominance of the global algorithm over the standard local-move methods, inherently due to the more efficient way of sampling the space of feasible solutions to the underdetermined system of linear equations underlying the structure of the problem. At each step of our algorithm the entire table, i.e. the entire

vector x , potentially changes, and not only by ± 1 . This generates sequences of values with relatively low correlations between iterates, and clearly facilitates an effective exploration of the multidimensional support in question.

The imputation of missing counts in two-way tables is illustrative of what is a promising approach to related problems of greater complexity, including tables of more than two dimensions and with more general patterns of missing data. In higher dimensional problems, with the cardinality of the space of feasible solutions increasing dramatically with dimension, the need for fast and statistically efficient methods of imputation are even more acute. Further work should explore variants of the global algorithm here and their extension to more complex models and problems.

7. APPENDIX: TECHNICAL NOTES

7.1. IRREDUCIBILITY OF THE GLOBAL MOVE CHAIN

We sketch the proof of convexity of the support of the uni-dimensional conditional posteriors distributions, and connectedness of the support of the joint posterior, for the vector x_2 . All notation is as used in Section 3. First, consider the uni-dimensional distribution of component i conditional on the remaining components of the vector, namely $p(x_{2,i}|x_{2,-a})$ for any $i = 1, \dots, (r-1)(c-1)$. We prove that the support is convex, as follows.

Suppose that two $(r-1)(c-1)$ -dimensional vectors u, v differ in just one of their components, i.e., $u = (u_1, \dots, u_i, \dots, u_{(r-1)(c-1)})'$ and $v = (u_1, \dots, v_i, \dots, u_{(r-1)(c-1)})'$. Suppose also that, for any vector y , the inequalities $A_1^{-1}(y - A_2u) \geq 0$ and $A_1^{-1}(y - A_2v) \geq 0$ hold. For any number $\alpha \in [0, 1]$ define the vector

$$z = \alpha u + (1 - \alpha)v = (u_1, \dots, \alpha u_i + (1 - \alpha)v_i, \dots, u_{(r-1)(c-1)})'.$$

Then z also satisfies $A_1^{-1}(y - A_2z) \geq 0$. In fact,

$$\begin{aligned} A_1^{-1}(y - A_2z) &= A_1^{-1}(y - A_2(\alpha u + (1 - \alpha)v)) \\ &= A_1^{-1}(\alpha y + (1 - \alpha)y - A_2(\alpha u + (1 - \alpha)v)) \\ &= A_1^{-1}\alpha(y - A_2u) + A_1^{-1}(1 - \alpha)(y - A_2v). \end{aligned}$$

This is the mean of two non-negative quantities by assumption, and so is itself a non-negative quantity. This ends the proof.

We now show that the support of the $(r-1)(c-1)$ -dimensional vector x_2 is fully connected. In fact, we show that its marginal subsets, which we just proved to be convex, are connected. The result implies irreducibility of the Markov chain and hence its convergence to the equilibrium distribution of interest. The proof is for the simplified case of a two-dimensional vector x_2 . Details for the general case can be found in Tebaldi and West (1998).

The proof is by contradiction: if we assume that both x_2 and $x_2 + (1, 1)'$ are feasible, either $x_2 + (1, 0)$ is feasible as well, or $x_2 + (0, 1)$ is, if not both. In fact, the two pairs of assumptions

$$\{A_1^{-1}(y - A_2x_2) \geq 0, \quad A_1^{-1}(y - A_2(x_2 + (1, 1)')) \geq 0\}$$

and

$$\{A_1^{-1}(y - A_2(x_2 + (1, 0)')) < 0, \quad A_1^{-1}(y - A_2(x_2 + (0, 1)')) < 0\}$$

are incompatible. To begin, note that $A_1^{-1}(y - A_2(x_2 + (1, 1)')) \geq 0$ reduces to

$$A_1^{-1}(y - A_2x_2 - (A_2^{11}, A_2^{21})' - (A_2^{12}, A_2^{22})') \geq 0$$

where the matrix elements are denoted by double superscript. Now add $A_1^{-1}(y - A_2x_2) \geq 0$ by assumption, to lead to

$$A_1^{-1}(y - A_2x_2) + A_1^{-1}(y - A_2x_2 - (A_2^{11}, A_2^{21})' - (A_2^{12}, A_2^{22})') \geq 0.$$

This reduces to

$$A_1^{-1}(y - A_2x_2 + y - A_2x_2 - (A_2^{11}, A_2^{21})' - (A_2^{12}, A_2^{22})') \geq 0$$

or

$$A_1^{-1}(y - A_2x_2 - (A_2^{11}, A_2^{21})' + y - A_2x_2 - (A_2^{12}, A_2^{22})') \geq 0.$$

This last term cannot then be the sum of two negative quantities, and this establishes the contradiction between the two pairs of inequalities stated at the beginning.

The same procedure works in cases when x_2 and $x_2 - (1, 1)'$ are feasible, proving that at least one of $x_2 - (1, 0)$ or $x_2 - (0, 1)$ is feasible, (or analogously, x_2 and $x_2 + (-1, 1)'$ or x_2 and $x_2 + (1, -1)'$ are).

This completes the proof for the two-dimensional case.

7.2. BOUNDS OF THE SUPPORT OF EACH x_{ij}

The use of a Metropolis-Hastings step in the generation of a new value for the single component of the vector x_2 automatically rejects inadmissible candidate values. The regions of admissibility are dictated by the values of the conditioning elements of the x_2 vector together with the positivity constraint on elements of x_1 . The computation of the resulting bounds before each simulation step is generally possible (see Tebaldi and West 1998, Section 3, for details) but can be a heavy computational burden in problems other than trivial. In the specific case of contingency table simulation, gross bounds can be relatively easily determined in advance by looking at the marginal sums, though we have not found the implementation of such an approach to significantly improve the already satisfactory rates of acceptance of our algorithm (consistently around 40% for the various components of the x vector).

7.3. STARTING VALUES FOR THE SIMULATION

The search for a feasible starting point for the simulation of the x vector components (i.e. the search for a set of cell counts satisfying the given marginal sums) can be made automatic by a simple manipulation of the structure of any table. See this as follows.

Arrange the given margins in decreasing order, and then fill in the cells of the permuted table by running through first the rows, from top to bottom, and columns, left to right, adding 1 to each element in turn so long as the table remains consistent with the marginal totals. The result is a table whose cell counts decrease from left to right and top to bottom. This table represents a feasible initial point for the iteration, after re-permuting the order of its cells to re-establish the original order of the marginal sums.

As an example, Table 1 can be initialised by reordering the column of the following table, obtained by reordering rows and columns as below, and then filling the cells by the above described algorithm.

	Hair color				
Eye color	Brunette	Blonde	Black	Red	Total
Brown	124	43	34	19	220
Blue	121	43	33	18	215
Hazel	25	25	25	18	93
Green	16	16	16	16	64
Total	286	127	108	71	592

REFERENCES

- AGRESTI, A. (1992) A survey of exact inference for contingency tables. *Statistical Science*, **7**, 131-177.
- BEST, N.G., COWLES, M.K., & VINES, S.K. (1995) *CODA: convergence diagnosis and output analysis software for Gibbs sampler output* (Version 0.3), Cambridge UK: Medical Research Council Biostatistics Unit.
- DIACONIS, P. and EFRON, B. (1985) Testing for independence in a two-way table: New interpretations of the Chi-square statistic. *Annals of Statistics*, **13**, 845-874.
- DIACONIS, P. and STURMFELS, B. (1998) Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, **26**, 1, 363-397.
- FIENBERG, S.E. (1997) Confidentiality and disclosure limitation methodology. *Technical Report*, Department of Statistics, Carnegie Mellon University.
- SMITH, P.W.F., FORSTER, J.J. and McDONALD, J.W. (1996) Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society*, (Ser. A), **159**, 2, 309-321.
- TEBALDI, C. and WEST, M. (1998) Bayesian inference on network traffic using link count data (with discussion). *Journal of the American Statistical Association*, **93**, 557-576.
- TIERNEY, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701-1762.
- WEST, M. (1994) Statistical inference for gravity models in transportation flow forecasting, *ISDS Discussion Paper #94-40*, Duke University, and *Technical Report #60*, National Institute of Statistical Sciences.

Figure 1 : Observed marginal counts for Block 9. I (family income) is on the left; W (number of workers in the family) is on the right.

Figure 2 : Observed marginal counts for Block 28. I (family income) is on the left; W (number of workers in the family) is on the right.

Figure 3 : Collateral table of full cell counts z used in analysis of Block 9 and 28 marginal totals. Darker shading corresponds to higher frequency.

Figure 4 : Observed marginal counts from the collateral table z .

Figure 5 : Marginal posterior distributions for cell probabilities underlying the table from Block 9. The histogram represents the marginal distributions for the π_{ij} in this case, and the approximate posterior means for the π_{ij} are noted above the histograms. The label "S" on the horizontal axis indicates the observed relative frequency in the corresponding cell of the collateral table z .

Figure 6 : Marginal posterior distributions for cell probabilities underlying the table from Block 28. The histogram represents the marginal distributions for the π_{ij} in this case, and the approximate posterior means for the π_{ij} are noted above the histograms. The label "S" on the horizontal axis indicates the observed relative frequency in the corresponding cell of the collateral table z .

Figure 7 : Three sets of marginal counts for Block 9, with distributions for (I in the left column, and for W in the right columns). The first row displays the observed margins of the collateral table z , the second row displays the actual observed margins of Block 9, and the third row gives the margins of the posterior predictive distribution for the full table for this block arising from the MCMC analysis.

Figure 8 : Three sets of marginal counts for Block 28, with distributions for (I in the left column, and for W in the right columns). The first row displays the observed margins of the collateral table z , the second row displays the actual observed margins of Block 28, and the third row gives the margins of the posterior predictive distribution for the full table for this block arising from the MCMC analysis.

Figure 9 : The sample autocorrelation functions of a randomly selected λ parameter computed from the output streams of two analyses: one (upper frame) using the standard local-move imputation algorithm, and the second (lower frame) using the new global method. The autocorrelations plotted are for values ten iterates apart, so that the Lag index represents ten steps through the algorithm.

Figure 10 : Posterior marginal distributions of the 16 cell counts underlying Table 1, based on 10,000 iterations of the global algorithm.

Figure 11 : Posterior marginal distributions of the 16 cell counts underlying Table 1, based on 10,000 iterations of the local-move algorithm.

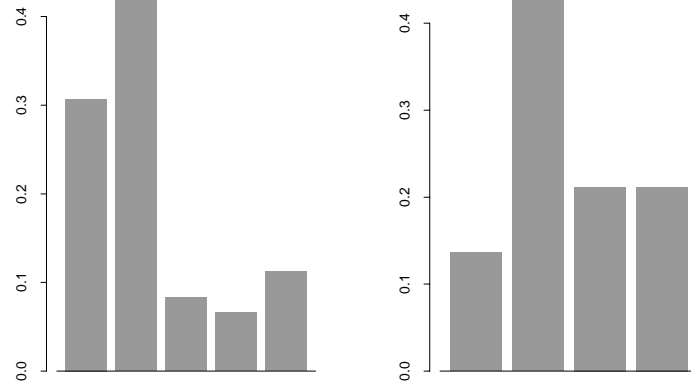


Figure 1:

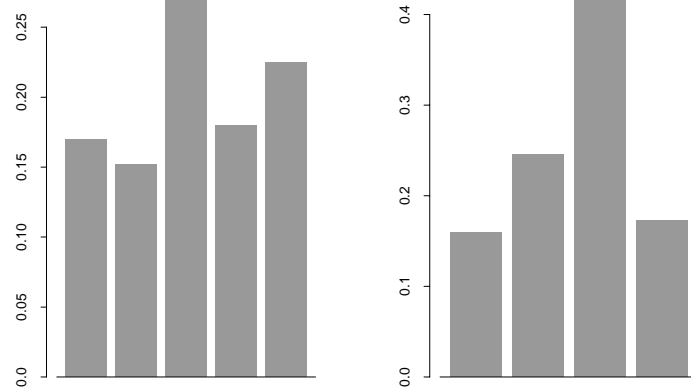


Figure 2:

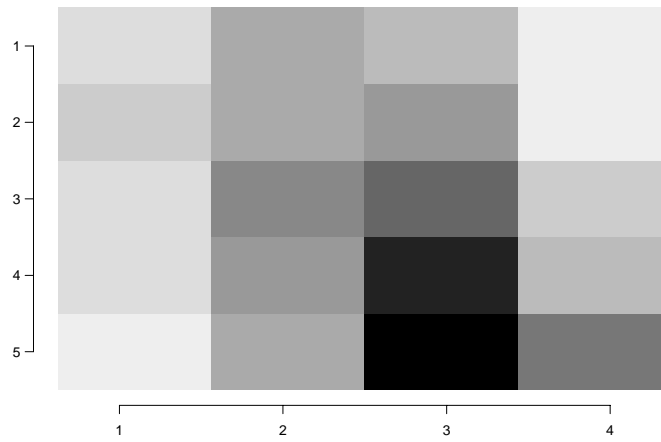


Figure 3:

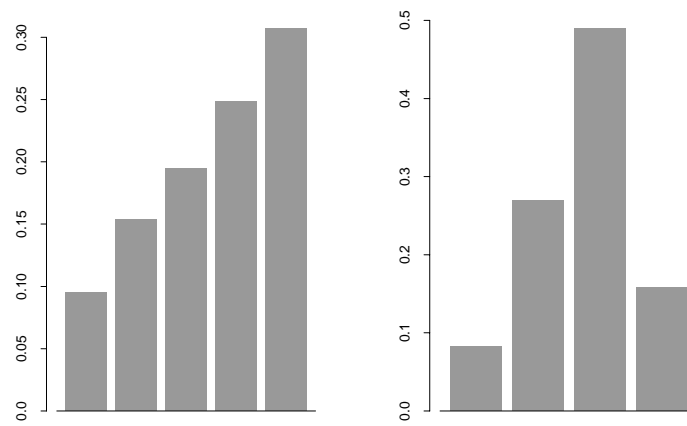


Figure 4:

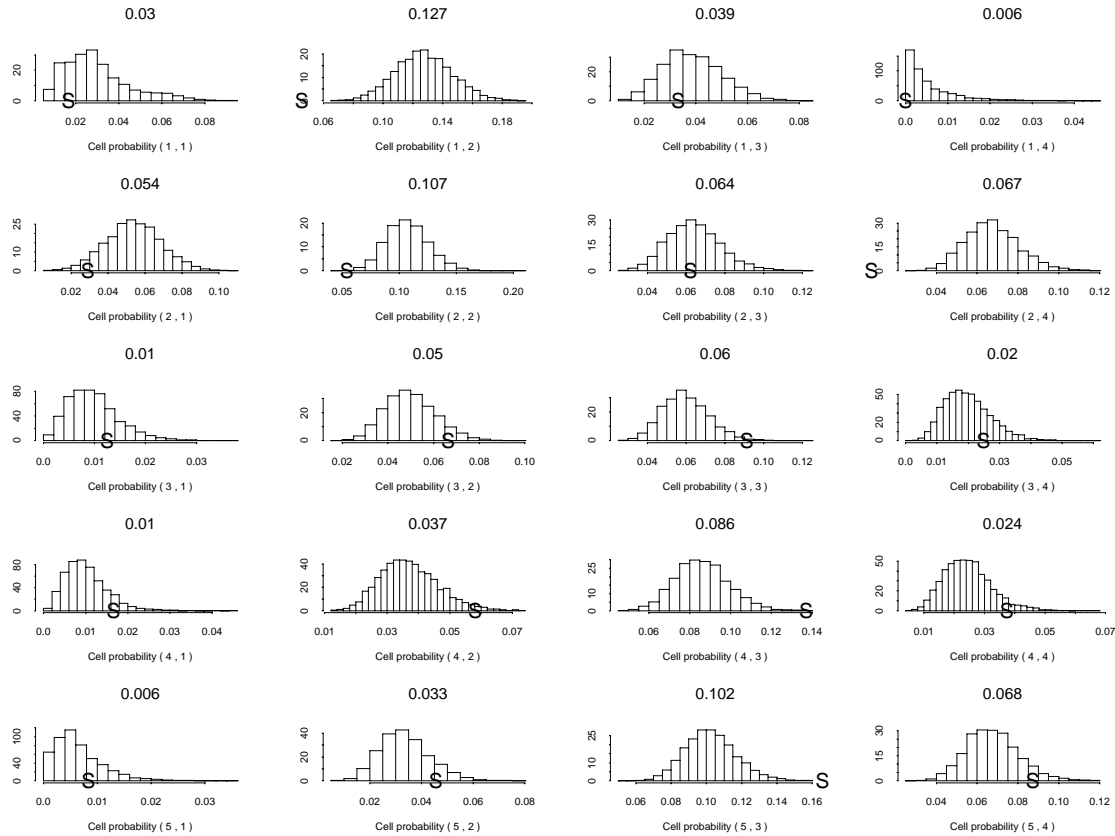


Figure 5:

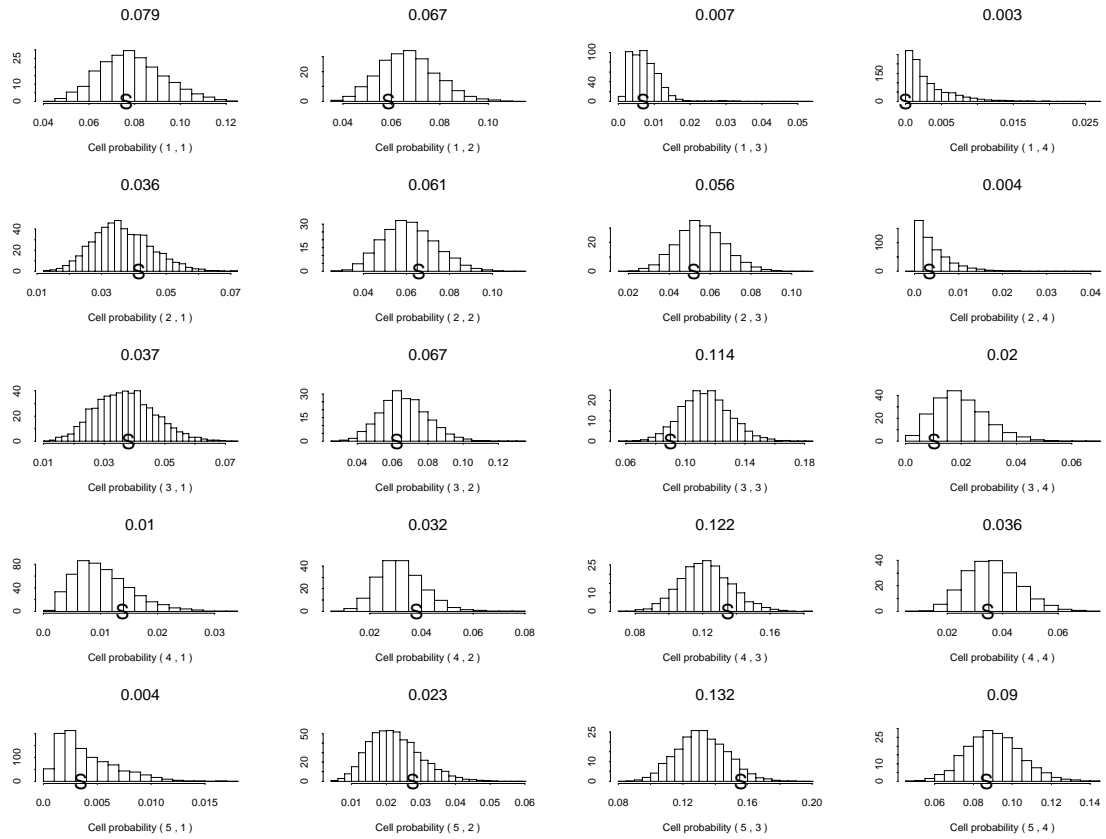


Figure 6:

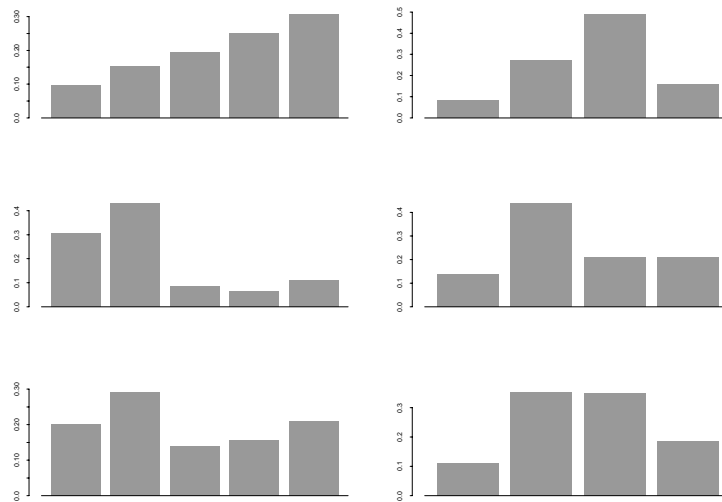


Figure 7:

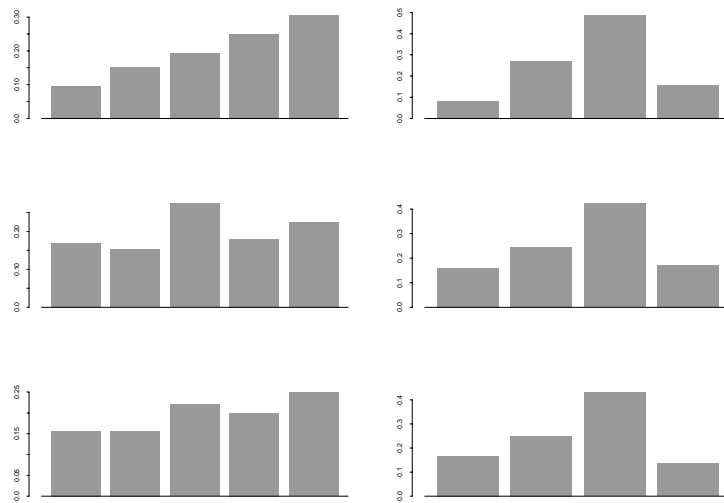
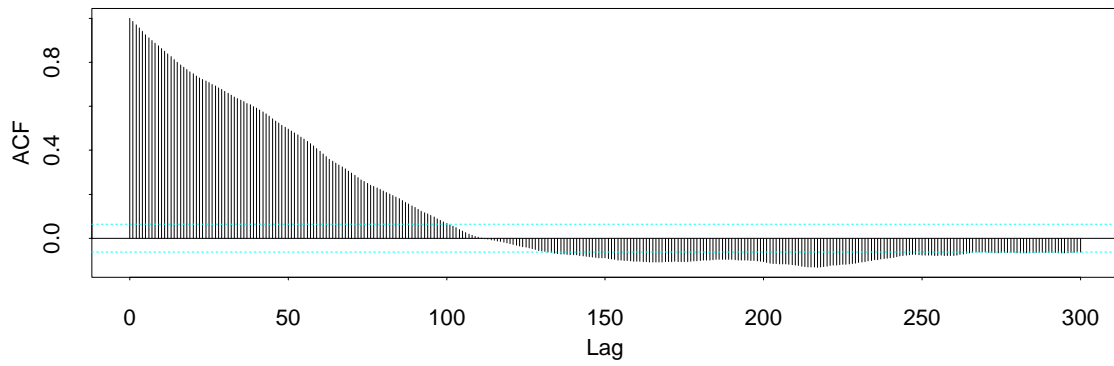


Figure 8:

Local move method



Global move method

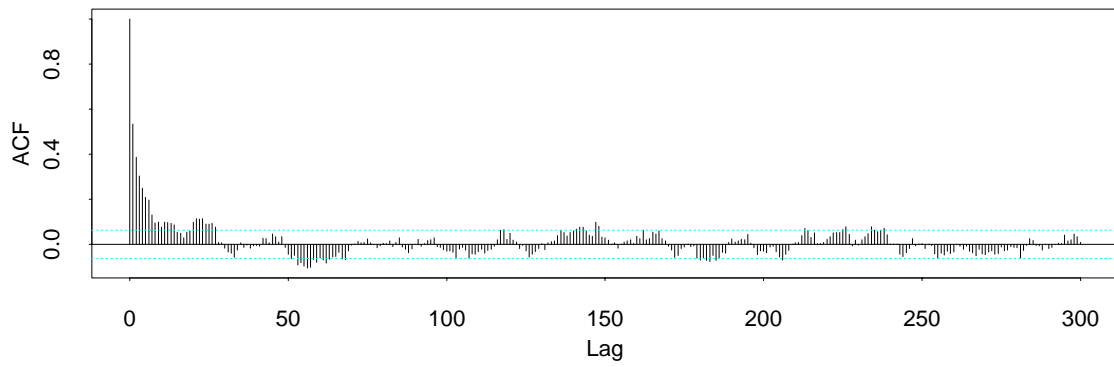


Figure 9:

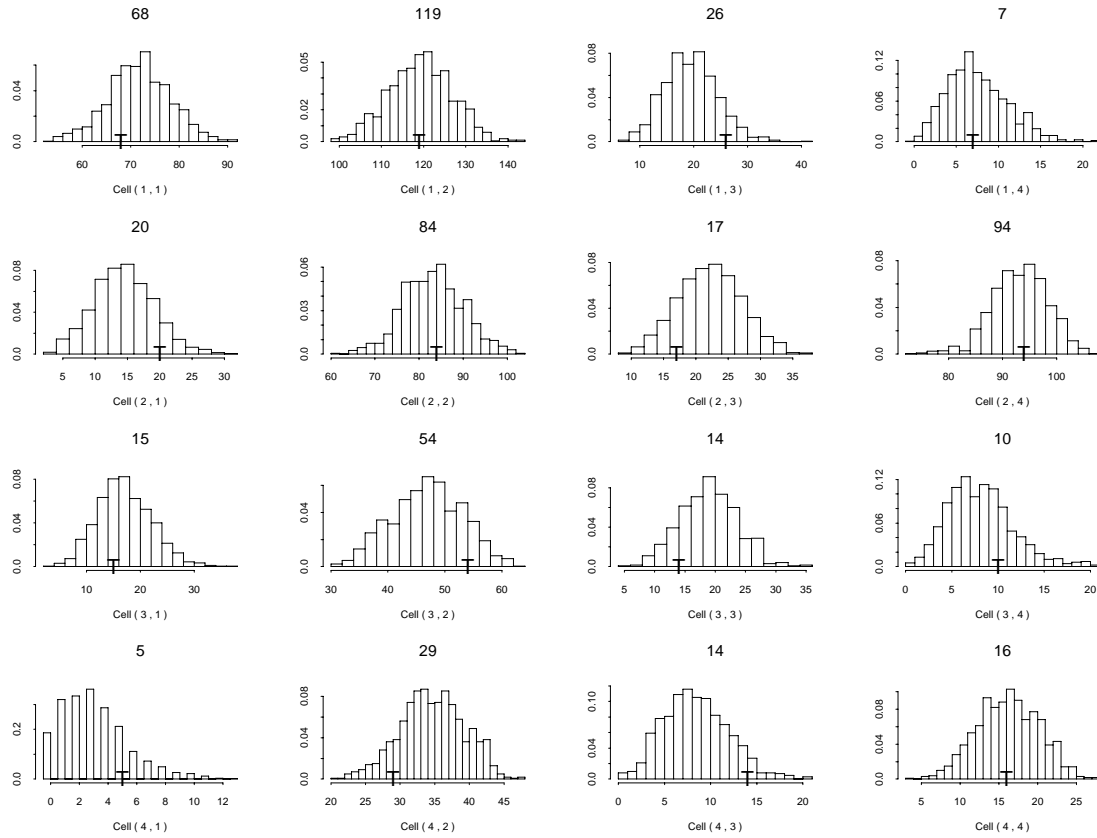


Figure 10:

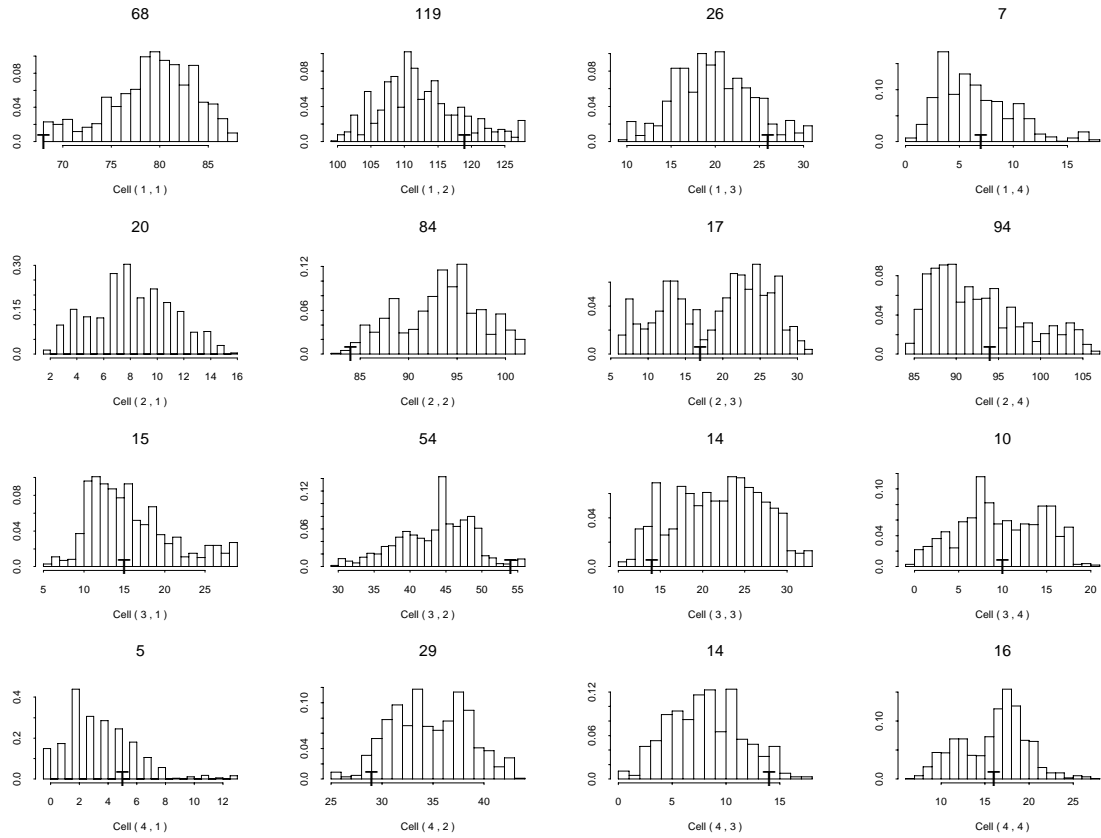


Figure 11: