The Generalized Shuttle Algorithm [1]

Adrian Dobra
University of Washington, Seattle

Stephen E. Fienberg
Carnegie Mellon University, Pittsburgh

Working Paper no. 83
Center for Statistics and the Social Sciences
University of Washington

February 29, 2008

[1]Adrian Dobra is Assistant Professor of Statistics and Nursing, Department of Statistics and Biobehavioral Nursing and Health Systems, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: adobra@stat.washington.edu; Web: www.stat.washington.edu/adobra. Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science, Department of Statistics, the Machine Learning Department and Cylab, Carnegie Mellon University, Pittsburgh PA 15213-3890. E-mail: fienberg@stat.cmu.edu; Web: www.stat.cmu.edu/∼fienberg/.

Abstract

Bounds for the cell counts in multi-way contingency tables given a set of marginal totals arise in a variety of different statistical contexts including disclosure limitation. We describe the Generalized Shuttle Algorithm for computing integer bounds of multi-way contingency tables induced by arbitrary linear constraints on cell counts. We study the convergence properties of our method by exploiting the theory of discrete graphical models and demonstrate the sharpness of the bounds for some specific settings. We give a procedure for adjusting these bounds to the sharp bounds that can also be employed to enumerate all tables consistent with the given constraints. Our algorithm for computing sharp bounds and enumerating multi-way contingency tables is the first approach that relies exclusively on the unique structure of the categorical data and does not employ any other optimization techniques such as linear or integer programming. We illustrate how our algorithm can be used to compute exact p-values of goodness-of-fit tests in exact conditional inference.

KEY WORDS: Bounds; Contingency tables; Exact testing; Log-linear models.

1 Introduction

Many statistical research problems involve working with sets of multi-way contingency tables defined by a set of constraints (e.g., marginal totals or structural zeroes). Four inter-related aspects involve: (1) the computation of sharp integer bounds, (2) counting, (3) exhaustive enumeration, and (4) sampling. Each of these areas or some combination of them play important roles in solving complex data analysis questions arising in seemingly unrelated fields. The computation of bounds is central to the task of assessing the disclosure risk of small cell counts (e.g., cells with entires of 1 or 2) when releasing marginals from a high-dimensional sparse contingency table — see, for example, Fienberg (1999) Dobra and Fienberg (2000); Dobra (2001). Another aspect of disclosure risk assessment involves counting feasible tables consistent with the release (see Fienberg and Slavkovic (2004, 2005)) or by estimating probability distributions on multi-way tables as in Dobra et al. (2003a).

Guo and Thompson (1992) employ sampling from a set of contingency tables to perform exact tests for Hardy-Weinberg proportions. Markov chain Monte Carlo (MCMC) sampling methods depend on the existence of a Markov basis that connects any two feasible tables through a series of Markov moves. Diaconis and Sturmfels (1998) were the first to show how to produce such moves through algebraic geometry techniques. Dobra (2003) gave formulas for Markov bases in the case of decomposable graphical models, while Dobra and Sullivant (2004) extend this work to reducible graphical models. Markov bases are local moves that change only a relatively small number of cell counts and can be contrasted with global moves that potentially alter all the counts. Dobra et al. (2006) describe how to produce global moves in a set of contingency tables by sequentially adjusting upper and lower bounds as more cells are fixed at certain values. Chen et al. (2006) present a similar method for finding feasible tables. Their sequential importance sampling approach seems to be more efficient than other MCMC techniques and builds on computational commutative algebra techniques to find bounds and to make random draws from the implied marginal cell distributions. Other work on algebraic geometry related to the theory of discrete graphical models includes Geiger et al. (2006)) and Hosten and Sturmfels (2007).

A special class of bounds we are interested in is attributed to Fréchet (1940), whose original presentation was in terms of cumulative distribution functions (c.d.f.) for a random vector $(D_1, D_2, \ldots, D_m)$ in R^m:

$$F_{1,2,\ldots,m}(x_1, x_2, \ldots, x_m) = \Pr(D_1 \leq x_1, D_2 \leq x_2, \ldots, D_m \leq x_m), \tag{1}$$

which are essentially equivalent to contingency tables when the underlying variables are

categorical. For example, suppose we have a two-dimensional table of counts, $\{n_{ij}\}$ adding up to the total $n_{++} = n$. If we normalize each entry by dividing by n and then create a table of partial sums, by cumulating the proportions from the first row and first column to the present ones, we have a set of values of the form (1). Thus, Fréchet bound results for distribution functions correspond to bounds for the cell counts where the values $\{x_i\}$ in (1) represent"cut-points" between categories for the ith categorical variable. Bonferroni (1936) and Hoeffding (1940) independently developed related results on bounds. Dobra and Fienberg (2000) proved that, when the fixed set of marginals defines a decomposable independence graph, the upper Fréchet bounds are the minimum of relevant margins, while the lower Fréchet bounds are the maximum of zero, or sum of the relevant margins minus the separators.

In this paper we propose an algorithm that we call the Generalized Shuttle Algorithm (GSA) which we can use to compute sharp integer bounds and exhaustively enumerate all feasible tables consistent with a set of constraints. Dobra et al. (2003b) provided a brief account of this work, while Dobra et al. (2006) showed its application to sampling contingency tables. Our procedure is deterministic and exploits the special structure of contingency tables, building on the work of Buzzigoli and Giusti (1999) who proposed the first version of the shuttle algorithm. Their innovative iterative approach simultaneously calculates bounds for all the cells in the table by sequentially alternating between upper and lower bounds; however, their version of the shuttle algorithm fails to converge to the sharp bounds for most configurations of fixed marginal totals, e.g., see Cox (1999). The explanation for this failure lies in the incomplete description of the dependencies among the cells of a contingency table used by Buzzigoli and Giusti. Chen et al. (2006) give an excellent discussion about the relationship between linear programming (LP), integer programming (IP) and the computation of bounds for contingency tables.

This paper is organized as follows. In Section 3 we give the basic definitions and notations. We present the full description of GSA in Section 4. In Sections 5 and 6, we describe two particular cases when the shuttle procedure converges to the sharp bounds. In Section 7 we present an approach for adjusting the shuttle bounds to the sharp bounds and also shows how to transform this procedure to enumerate multi-way tables. In Section 8 we show that GSA is able to efficiently compute bounds for a sixteen-way sparse contingency table. In Section 9 we give six examples that illustrate how GSA can be used for computing bounds as well as exact p-values based on the hypergeometric distribution. Complete proofs of our theoretical results are given in the Appendix. Source code implementing GSA is available for download from

`http://www.stat.washington.edu/adobra/software/gsa/`

2 Terminology and Notation

Let $X = (X_1, X_2, \ldots, X_k)$ be a vector of k discrete random variables cross-classified in a frequency count table $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$, where $\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2 \times \ldots \times \mathcal{I}_k$ and X_r takes the values $\mathcal{I}_r := \{1, 2, \ldots, I_r\}$. Denote $K = \{1, 2, \ldots, k\}$. For $r \in K$, we denote by $\mathcal{P}(\mathcal{I}_r)$ the set of all partitions of $\mathcal{I}_r = \mathcal{I}_r^1 \cup \mathcal{I}_r^2 \cup \ldots \cup \mathcal{I}_r^{l_r}$, i.e.,

$$\begin{aligned} \mathcal{P}(\mathcal{I}_r) \;:=\; \Big\{ \big\{ \mathcal{I}_r^1, \mathcal{I}_r^2, \ldots, \mathcal{I}_r^{l_r} \big\} : 1 \le l_r \le I_r;\; \mathcal{I}_r^{l^1} \cap \mathcal{I}_r^{l^2} = \emptyset, \\ \text{if } 1 \le l^1 \ne l^2 \le l_r,\; \mathcal{I}_r^{l^1} \subset \mathcal{I}_r, \text{ for } 1 \le l^1 \le l_r \Big\}. \end{aligned}$$

Let $\mathcal{RD}$ be the set of marginal tables obtainable by aggregating $\mathbf{n}$ not only across variables, but also across categories within variables. We can uniquely determine a table $\mathbf{n}' \in \mathcal{RD}$ from $\mathbf{n}$ by choosing $\mathcal{I}_1' \in \mathcal{P}(\mathcal{I}_1)$, $\mathcal{I}_2' \in \mathcal{P}(\mathcal{I}_2)$, ..., $\mathcal{I}_k' \in \mathcal{P}(\mathcal{I}_k)$. We write

$$\mathbf{n}' = \{ n'(J_1, J_2, \ldots, J_k) : (J_1, J_2, \ldots, J_k) \in \mathcal{I}_1' \times \mathcal{I}_2' \times \ldots \times \mathcal{I}_k' \}, \tag{2}$$

where the entries of $\mathbf{n}'$ are sums of appropriate entries of $\mathbf{n}$:

$$n'(J_1, J_2, \ldots, J_k) := \sum_{i_1 \in J_1} \sum_{i_2 \in J_2} \cdots \sum_{i_k \in J_k} n_K(i_1, i_2, \ldots, i_k).$$

We associate the table $\mathbf{n}$ with $\mathcal{I}_r' = \{\{1\}, \{2\}, \ldots, \{I_r\}\}$, for $r = 1, \ldots, k$. Choosing $\mathcal{I}_r' = \{\mathcal{I}_r\}$ is equivalent to collapsing across the r-th variable. The *dimension* of $\mathbf{n}' \in \mathcal{RD}$ is the number of variables cross-classified in $\mathbf{n}'$ that have more than one category. We obtain the C-marginal $\mathbf{n}_C$ of $\mathbf{n}$, $C \subset K$, by taking

$$\mathcal{I}_r' = \begin{cases} \{\{1\}, \{2\}, \ldots, \{I_r\}\}, & \text{if } r \in C, \\ \mathcal{I}_r, & \text{otherwise,} \end{cases}$$

for $r = 1, 2, \ldots, k$. The dimension of $\mathbf{n}_C$ is equal to the number of elements in C. The grand total of $\mathbf{n}$ has dimension zero, while $\mathbf{n}$ has dimension k.

We introduce the set of tables $\mathcal{RD}(\mathbf{n}')$ containing the tables $\mathbf{n}'' \in \mathcal{RD}$ obtainable from $\mathbf{n}'$ by table redesign such that $\mathbf{n}''$ and $\mathbf{n}'$ have the same dimension. We have $\mathbf{n}' \in \mathcal{RD}(\mathbf{n}')$ and $\mathcal{RD}(n_\emptyset) = \{n_\emptyset\}$, where $n_\emptyset$ is the grand total of $\mathbf{n}$. The set $\mathcal{RD}$ itself results from aggregating every marginal $\mathbf{n}_C$ of $\mathbf{n}$ across categories, such that every variable having at least two categories in $\mathbf{n}_C$ also has at least two categories in the new redesigned table:

$$\mathcal{RD} = \bigcup \{\mathcal{RD}(\mathbf{n}_C) : C \subseteq K\}. \tag{3}$$

We formally define $\mathbf{T}$ as

$$\mathbf{T} := \{t_{J_1 J_2 \ldots J_k} : \emptyset \neq J_1 \times J_2 \times \ldots \times J_k \subseteq \mathcal{I}_1 \times \mathcal{I}_2 \times \ldots \times \mathcal{I}_k\}, \tag{4}$$

where

$$t_{J_1 J_2 \ldots J_k} = \sum_{i_1 \in J_1} \sum_{i_2 \in J_2} \ldots \sum_{i_k \in J_k} n_K(i_1, i_2, \ldots, i_k). \tag{5}$$

The elements in $\mathbf{T}$ are blocks or "super-cells" formed by joining table entries in $\mathbf{n}$. These blocks can be viewed as entries in a k-dimensional table that cross-classifies the variables $(Y_j : j = 1, 2, \ldots, k)$, where Y_j takes values $y_j \in \left\{\mathcal{I}'_j : \emptyset \neq \mathcal{I}'_j \subseteq \mathcal{I}_j\right\}$. The number of elements in $\mathbf{T}$ is $\#(\mathbf{T}) = \prod_{r=1}^{k} \left(2^{I_r} - 1\right)$.

If the set of cell entries in $\mathbf{n}$ that define a "super-cell" $t_2 = t_{J_1^2 \ldots J_k^2} \in \mathbf{T}$ includes the set of cells defining another "super-cell" $t_1 = t_{J_1^1 \ldots J_k^1} \in \mathbf{T}$, we write $t_1 = t_{J_1^1 \ldots J_k^1} \prec t_2 = t_{J_1^2 \ldots J_k^2}$. We formally define the partial ordering "$\prec$" on the cells in $\mathbf{T}$ by

$$t_{J_1^1 J_2^1 \ldots J_k^1} \prec t_{J_1^2 J_2^2 \ldots J_k^2} \Leftrightarrow J_1^1 \subseteq J_1^2, J_2^1 \subseteq J_2^2, \ldots, J_k^1 \subseteq J_k^2.$$

This partial ordering, $(\mathbf{T}, \prec)$, has a maximal element, namely the grand total $n_\emptyset = t_{\mathcal{I}_1 \mathcal{I}_2 \ldots \mathcal{I}_k}$ of the table and several minimal elements — the actual cell counts $n(i) = n(i_1, i_2, \ldots, i_k) = t_{\{i_1\}\{i_2\}\ldots\{i_k\}}$. Thus, we can represent he lattice, $(\mathbf{T}, \prec)$, as a hierarchy with the grand total at the top level and the cells counts $n(i)$ at the bottom level. If $t_1 = t_{J_1^1 J_2^1 \ldots J_k^1}$ and $t_2 = t_{J_1^2 J_2^2 \ldots J_k^2}$ are such that $t_1 \prec t_2$ with $J_r^1 = J_r^2$, for $r = 1, \ldots, r_0 - 1, r_0 + 1, \ldots, k$ and $J_{r_0}^1 \neq J_{r_0}^2$, we define the *complement* of the cell t_1 with respect to t_2 to be the cell $t_3 = t_{J_1^3 J_2^3 \ldots J_k^3}$, where

$$J_r^3 = \begin{cases} J_r^1, & \text{if } r \neq r_0, \\ J_r^2 \setminus J_r^1, & \text{if } r = r_0, \end{cases},$$

for $r = 1, 2, \ldots, k$. We write $t_1 \oplus t_3 = t_2$. The elements in $\mathbf{T}$ are blocks formed by joining table entries in $\mathbf{n}$. The operator "$\oplus$" is equivalent to joining two blocks of cells in $\mathbf{T}$ to form a third block where the blocks to be joined have the same categories in $(k-1)$ dimensions and they cannot to share any categories in the remaining dimension.

3 The Generalized Shuttle Algorithm

The fundamental idea behind the Generalized Shuttle Algorithm (GSA) is that the upper and lower bounds for the cells in $\mathbf{T}$ are interlinked, i.e., bounds for some cells in $\mathbf{T}$ induce

bounds for some other cells in $\mathbf{T}$. We can improve (tighten) the bounds for all the cells in which we are interested until we can make no further adjustment. Although Buzzigoli and Giusti (1999) sketched this innovative idea, they did not accurately identify or exploit the special hierarchical structure of $\mathbf{T}$.

Let $L_0(\mathbf{T}) := \{L_0(t) : t \in \mathbf{T}\}$ and $U_0(\mathbf{T}) := \{U_0(t) : t \in \mathbf{T}\}$ be initial upper and lower bounds. By default we set $L_0(t) = 0$ and $U_0(t) = n_\emptyset$, but we can express almost any type of information about the counts in cells $\mathbf{T}$ using these bounds. For example, a known count c in a cell t with a fixed marginal implies that $L_0(t) = U_0(t) = c$. A cell t that can take only two values 0 or 1 has $L_0(t) = 0$ and $U_0(t) = 1$.

We denote by $S[L_0(\mathbf{T}), U_0(\mathbf{T})]$ the set of integer feasible arrays $V(\mathbf{T}) := \{V(t) : t \in \mathbf{T}\}$ consistent with $L_0(\mathbf{T})$ and $U_0(\mathbf{T})$: (i) $L_0(t) \leq V(t) \leq U_0(t)$, for all $t \in \mathbf{T}$, and (ii) $V(t_1) + V(t_3) = V(t_2)$, for all $(t_1, t_2, t_3) \in \mathcal{Q}(\mathbf{T})$, where

$$\mathcal{Q}(\mathbf{T}) := \{(t_1, t_2, t_3) \in \mathbf{T} \times \mathbf{T} \times \mathbf{T} : t_1 \oplus t_3 = t_2\}. \tag{6}$$

We let $\mathcal{N} \subset \mathbf{T}$ be the set of cells in table $\mathbf{n}$. A feasible table consistent with the constraints imposed (e.g., fixed marginals) is $\{V(t) : t \in \mathcal{N}\}$ where $V(\mathbf{T}) \in S[L_0(\mathbf{T}), U_0(\mathbf{T})]$.

The sharp integer bounds $[L(t), U(t)]$, $t \in \mathbf{T}$, are the solution of the integer optimization problems:

$$\min\{\pm V(t) : V(\mathbf{T}) \in S[L_0(\mathbf{T}), U_0(\mathbf{T})]\}. \tag{7}$$

We initially set $L(\mathbf{T}) = L_0(\mathbf{T})$ and $U(\mathbf{T}) = U_0(\mathbf{T})$, and sequentially improve these loose bounds by GSA until we get convergence. Consider $\mathbf{T}_0 := \{t \in \mathbf{T} : L(t) = U(t)\}$ to be the cells with the current lower and upper bounds equal. We say that the remaining cells in $\mathbf{T} \setminus \mathbf{T}_0$ are *free*. As the algorithm progresses, we improve the bounds for the cells in $\mathbf{T}$ and add more and more cells to $\mathbf{T}_0$. For each t in $\mathbf{T}_0$, we assign a value $V(t) := L(t) = U(t)$.

We sequentially go through the dependencies $\mathcal{Q}(\mathbf{T})$ and update the upper and lower bounds in the following fashion. Consider a triplet $(t_1, t_2, t_3) \in \mathcal{Q}(\mathbf{T})$. We have $t_1 \prec t_2$ and $t_3 \prec t_2$. We update the upper and lower bounds of t_1, t_2 and t_3 so that the new bounds satisfy the dependency $t_1 \oplus t_3 = t_2$.

If all three cells have fixed values, i.e., $t_1, t_2, t_3 \in \mathbf{T}_0$, we check whether $V(t_1) + V(t_3) = V(t_2)$. If this equality does not hold, we stop GSA because $S[L_0(\mathbf{T}), U_0(\mathbf{T})]$ is empty – there is no integer table consistent with the constraints imposed.

Now assume that $t_1, t_3 \in \mathbf{T}_0$, and $t_2 \notin \mathbf{T}_0$. Then t_2 can take only one value, namely $V(t_1) + V(t_3)$. If $V(t_1) + V(t_3) \notin [L(t_2), U(t_2)]$, we encounter an inconsistency and stop. Otherwise we set $V(t_2) = L(t_2) = U(t_2) := V(t_1) + V(t_3)$ and include t_2 in $\mathbf{T}_0$. Similarly, if $t_1, t_2 \in \mathbf{T}_0$ and $t_3 \notin \mathbf{T}_0$, t_3 can only be equal to $V(t_2) - V(t_1)$. If $V(t_2) - V(t_1) \notin [L(t_3), U(t_3)]$,

we again discover an inconsistency. If this is not true, we set $V(t_3) = L(t_3) = U(t_3) := V(t_2) - V(t_1)$ and $\mathbf{T}_0 := \mathbf{T}_0 \cup \{t_3\}$. In the case when $t_2, t_3 \in \mathbf{T}_0$ and $t_1 \notin \mathbf{T}_0$, we proceed in an analogous manner.

Next we examine the situation when at least two of the cells t_1, t_2, t_3 do not have a fixed value. Suppose $t_1 \notin \mathbf{T}_0$. The new bounds for t_1 are

$$\begin{aligned} U(t_1) &:= \min\{U(t_1), U(t_2) - L(t_3)\}, \\ L(t_1) &:= \max\{L(t_1), L(t_2) - U(t_3)\}. \end{aligned}$$

If $t_3 \notin \mathbf{T}_0$, we update $L(t_3)$ and $U(t_3)$ in the same way. Finally, if $t_2 \notin \mathbf{T}_0$, we set

$$\begin{aligned} U(t_2) &:= \min\{U(t_2), U(t_1) + U(t_3)\}, \\ L(t_2) &:= \max\{L(t_2), L(t_1) + L(t_3)\}. \end{aligned}$$

After updating the bounds of some cell $t \in \mathbf{T}$, we check whether the new upper bound equals the new lower bound. If this is true we set $V(t) := L(t) = U(t)$ and include t in $\mathbf{T}_0$.

We continue iterating through all the dependencies in $\mathcal{Q}(\mathbf{T})$ until the upper bounds no longer decrease, the lower bounds no longer increase, and no new cells are added to $\mathbf{T}_0$. Therefore the procedure comes to an end if and only if we detect an inconsistency or if we cannot improve the bounds. One of these two events eventually occurs; hence the algorithm stops after a finite number of steps.

If we do not encounter any inconsistencies, the algorithm converges to bounds $L_s(\mathbf{T})$ and $U_s(\mathbf{T})$ that are not necessarily sharp: $L_s(t) \leq L_0(t) \leq U_0(t) \leq U_s(t)$. These arrays define the same feasible set of tables as the arrays $L_0(\mathbf{T})$ and $U_0(\mathbf{T})$ we started with, i.e., $S[L_s(\mathbf{T}), U_s(\mathbf{T})] = S[L_0(\mathbf{T}), U_0(\mathbf{T})]$, since the dependencies $\mathcal{Q}(\mathbf{T})$ need to be satisfied.

There exist two particular cases when we can easily prove that GSA converges to sharp integer bounds: (i) the case of a dichotomous k-dimensional table with all $(k-1)$-dimensional marginals fixed, and (ii) the case when the marginals we fix are the minimal sufficient statistics of a decomposable log-linear model. In both instances explicit formulas for the bounds exist. Employing GSA turns out to be equivalent to calculating the bounds directly as we prove in the next two sections.

4 Computing Bounds for Dichotomous k-way Cross-classifications Given All $(k-1)$-dimensional Marginals

Consider a k-way table $\mathbf{n} := \{n(i)\}_{i \in \mathcal{I}}$ with $\mathcal{I}_1 = \mathcal{I}_2 = \ldots = \mathcal{I}_k = \{1, 2\}$. The set $\mathbf{T}$ associated with $\mathbf{n}$ is the set of cells of every marginal of $\mathbf{n}$, while the set $\mathbf{T}_0$ of cells having a

fixed value is $\mathbf{T}_0 = \{n_C(i_C) : i_C \in \mathcal{I}_C \text{ for some } C \subset K,\ C \neq K\}$. The only cells in $\mathbf{T}$ that are not fixed are the cells in $\mathbf{n}$: $\mathbf{T} \setminus \mathbf{T}_0 = \{n(i) : i \in \mathcal{I}\}$.

The $(k-1)$-dimensional marginals of $\mathbf{n}$ are the minimal sufficient statistics of the log-linear model of no $(k-1)$-order interaction. Fienberg (1999) pointed that this log-linear model has only one degree of freedom because $\mathbf{n}$ is dichotomous, hence we can uniquely express the count in any cell $n(i)$, $i \in \mathcal{I}$, as a function of one single fixed cell alone.

Let n^* be the unknown count in the $(1, 1, \ldots, 1)$ cell. In Proposition 4.1 we give an explicit formula for computing the count in an arbitrary cell $n(i^0)$, $i^0 \in \mathcal{I}$, based on n^* and on the set of fixed marginals.

Proposition 4.1 *Let n^* be the count in the $(1, 1, \ldots, 1)$ cell. Consider an index $i^0 = (i^0_1, i^0_2, \ldots, i^0_k) \in \mathcal{I}$. Let $\{q_1, q_2, \ldots, q_l\} \subset K$ such that, for $r \in K$, we have*

$$i^0_r = \begin{cases} 1, & \text{if } \ r \in K \setminus \{q_1, q_2, \ldots, q_l\}, \\ 2, & \text{if } \ \ r \in \{q_1, q_2, \ldots, q_l\}. \end{cases} \tag{8}$$

For $s = 1, 2, \ldots, l$, denote $C_s := K \setminus \{q_s\}$. Then

$$n(i^0) = (-1)^l \cdot n^* - \sum_{s=0}^{l-1} (-1)^{l+s} \cdot n_{C_{(l-s)}}(1, \ldots, 1, i^0_{q_{(l-s)}+1}, \ldots, i^0_k). \tag{9}$$

We obtain the upper and lower bounds induced on the $(1, 1, \ldots, 1)$ cell count in table $\mathbf{n}$ by fixing the set of cells $\mathbf{T}_0$ by imposing the non-negativity constraints in equation (9). More explicitly, $n(i^0) \geq 0$ implies that the sharp lower bound for the count n^* is

$$L(n^*) = \max \left\{ \sum_{s=0}^{l-1} (-1)^s \cdot n_{C_{(l-s)}}(1, \ldots, 1, i^0_{q_{(l-s)}+1}, \ldots, i^0_k) : l \text{ even} \right\}, \tag{10}$$

whereas the sharp upper bound for the count n^* satisfies

$$U(n^*) = \min \left\{ \sum_{s=0}^{l-1} (-1)^s \cdot n_{C_{(l-s)}}(1, \ldots, 1, i^0_{q_{(l-s)}+1}, \ldots, i^0_k) : l \text{ odd} \right\}. \tag{11}$$

We are now ready to give the main result of this section:

Proposition 4.2 *The generalized shuttle algorithm converges to the bounds in equations (10) and (11).*

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

Table 1: Prognostic factors for coronary heart disease as measured on Czech autoworkers from Edwards and Havranek (1985).

4.1 Example: Bounds for the Czech Autoworkers data

Table 1 contains a 2^6 table, originally analyzed by Edwards and Havranek (1985) that cross-classifies binary risk factors denoted by A, B, C, D, E, F for coronary trombosis from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory. Here A indicates whether or not the worker "smokes", B corresponds to "strenuous mental work", C corresponds to "strenuous physical work", D corresponds to "systolic blood pressure", E corresponds to "ratio of β and α lipoproteins" and F represents "family anamnesis of coronary heart disease". We use GSA to calculate the bounds induced by fixing the five-way marginals – see Table 2. There are only two tables having this set of marginals. The second feasible table is obtained by adding or substracting one unit from the corresponding entries in Table 1.

5 Calculating Bounds in the Decomposable Case

Suppose we have p possibly overlapping marginal tables $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ such that $C_1 \cup C_2 \cup \ldots \cup C_p = K$. The index sets $\mathcal{C}(\mathcal{G}) := \{C_1, C_2, \ldots, C_p\}$ define a decomposable independence graph $\mathcal{G}$. Denote by $\mathcal{S}(\mathcal{G}) := \{S_2, \ldots, S_p\}$ the set of separators of the graph $\mathcal{G} = (K, E)$.

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		[44,45]	[39,40]	[111,112]	[67,68]
			yes		[128,129]	[145,146]	[12,13]	[22,23]
		≥ 140	no		[34,35]	[12,13]	[80,81]	[32,33]
			yes		[109,110]	[66,67]	[6,7]	[9,10]
	≥ 3	< 140	no		[22,23]	[32,33]	[70,71]	[65,66]
			yes		[50,51]	[79,80]	[6,7]	[13,14]
		≥ 140	no		[24,25]	[24,25]	[72,73]	[57,58]
			yes		[50,51]	[63,64]	[7,8]	[15,16]
pos	< 3	< 140	no		[4,5]	[7,8]	[21,22]	[8,9]
			yes		[9,10]	[16,17]	[0,1]	[4,5]
		≥ 140	no		[4,5]	[2,3]	[10,11]	[8,9]
			yes		[13,14]	[17,18]	[5,6]	[1,2]
	≥ 3	< 140	no		[7,8]	[2,3]	[13,14]	[14,15]
			yes		[8,9]	[16,17]	[2,3]	[2,3]
		≥ 140	no		[3,4]	[0,1]	[13,14]	[10,11]
			yes		[5,6]	[13,14]	[3,4]	[4,5]

Table 2: Bounds for entries in Table 1 induced by fixing the five-way marginals.

Every S_j is included in some clique C_i, hence the marginals $\mathbf{n}_{S_2}, \ldots, \mathbf{n}_{S_p}$ will also be fixed. The Fréchet bounds induced by this set of marginals are given by the following result due to Dobra and Fienberg (2000) and Dobra (2001):

Theorem 5.1

$$\begin{aligned} &\min\left\{n_{C_1}\left(i_{C_1}\right), \ldots, n_{C_p}\left(i_{C_p}\right)\right\} \\ &\geq n(i) \geq \max\left\{\sum_{j=1}^{p} n_{C_j}\left(i_{C_j}\right) - \sum_{j=2}^{p} n_{S_j}\left(i_{S_j}\right), 0\right\} \end{aligned} \tag{12}$$

are sharp bounds given the marginals $\mathbf{n}_{C_1}$, ..., $\mathbf{n}_{C_p}$.

We can derive analogous Fréchet bounds for each cell in the set of cells $\mathbf{T} = \mathbf{T}^{(\mathbf{n})}$ associated with table $\mathbf{n}$. First we attempt to develop inequalities for the cells contained in the marginals of $\mathbf{n}$:

$$\left\{n_D(i_D) : i_D \in \mathcal{I}_D \text{ for some } D \subset K\right\}.$$

Proposition 5.1 *For a subset* $D_0 \subset K$ *and an index* $i^0_{D_0} \in \mathcal{I}_{D_0}$, *the following inequalities*

hold:

$$\begin{aligned}
&\min\left\{n_{C\cap D_0}\left(i^0_{C\cap D_0}\right)|C\in\mathcal{C}(\mathcal{G})\right\}\\
&\geq n_{D_0}(i^0_{D_0})\\
&\geq \max\left\{0,\sum_{C\in\mathcal{C}(\mathcal{G})} n_{C\cap D_0}\left(i^0_{C\cap D_0}\right)-\sum_{S\in\mathcal{S}(\mathcal{G})} n_{S\cap D_0}\left(i^0_{S\cap D_0}\right)\right\}. \qquad (13)
\end{aligned}$$

The upper and lower bounds in equation (13) are defined to be the Fréchet bounds for the cell entry $n_{D_0}(i^0_{D_0})$ *given* $\mathbf{n}_{C_1},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_p}$.

For $D_0=K$, equation (13) becomes equation (12). At this point we know how to write Fréchet bounds for cell entries in an arbitrary table $\mathbf{n}'\in\mathcal{RD}$. If $\mathbf{n}'$ is not a proper marginal of $\mathbf{n}$, i.e., $\mathbf{n}'\notin\{\mathbf{n}_D : D\subset K\}$, from equation (3) we deduce that there exists $D_0\subset K$ such that $\mathbf{n}'\in\mathcal{RD}(\mathbf{n}_{D_0})$. Since the set of fixed marginals $\mathbf{n}_{C_1\cap D_0}$, $\mathbf{n}_{C_2\cap D_0}$, ..., $\mathbf{n}_{C_p\cap D_0}$ of $\mathbf{n}_{D_0}$ induce a decomposable independence graph $\mathcal{G}(D_0)$, we obtain $\mathbf{n}'$ from $\mathbf{n}_{D_0}$ by sequentially joining categories associated with the variables cross-classified in $\mathbf{n}_{D_0}$. If we apply exactly the same sequence of "join" operations to every marginal $\mathbf{n}_{C_r\cap D_0}$, $r=1,2,\ldots,p$, we end up with p fixed marginals $\mathbf{n}'_{C_1\cap D_0}$, $\mathbf{n}'_{C_2\cap D_0}$, ..., $\mathbf{n}'_{C_p\cap D_0}$ of $\mathbf{n}'$. The independence graph induced by those marginals coincides with $\mathcal{G}(D_0)$. Therefore the Fréchet bounds for a cell entry in $\mathbf{n}'$ are given either by Proposition 5.1 or by Theorem 5.1 if $\mathbf{n}'\in\mathcal{RD}(\mathbf{n})$.

The following lemma tells us that the Fréchet bounds for a cell $n_{D_0}(i^0_{D_0})$, $D_0\subset K$, are sharp if $\mathbf{n}$ has two fixed non-overlapping marginals.

Lemma 5.1 *Let* $\mathcal{G}=(K,E)$ *be a decomposable independence graph induced by the marginals* $\mathbf{n}_{C_1},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_p}$. *Consider a subset* $D_0\subset K$ *and let* $v\in K\setminus D_0$ *be a simplicial vertex of* $\mathcal{G}$. *It is known that a simplicial vertex belongs to precisely one clique, say* $v\in C_1$. *Then finding bounds for a cell* $n_{D_0}(i^0_{D_0})$, $i^0_{D_0}\in\mathcal{I}_{D_0}$, *given* $\mathbf{n}_{C_1},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_p}$ *is equivalent to finding bounds for* $n_{D_0}(i^0_{D_0})$ *given* $\mathbf{n}_{C_1\setminus\{v\}},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_p}$.

The Fréchet bounds for cells in a marginal $\mathbf{n}_{D_0}$ of $\mathbf{n}$ might not be the best bounds possible.

Lemma 5.2 *Assume there are two fixed marginals* $\mathbf{n}_{C_1}$ *and* $\mathbf{n}_{C_2}$ *such that* $C_1\cup C_2=K$, *but* $C_1\cap C_2=\emptyset$. *Consider* $D_0\subset K$. *The Fréchet bounds for* $n_{D_0}(i^0_{D_0})$ *given* $\mathbf{n}_{C_1}$ *and* $\mathbf{n}_{C_2}$

$$\begin{aligned}
&\min\left\{n_{C_1\cap D_0}(i^0_{C_1\cap D_0}),n_{C_2\cap D_0}(i^0_{C_2\cap D_0})\right\}\\
&\geq n_{D_0}(i^0_{D_0})\\
&\geq \max\left\{0,n_{C_1\cap D_0}(i^0_{C_1\cap D_0})+n_{C_2\cap D_0}(i^0_{C_2\cap D_0})-n_\emptyset\right\} \qquad (14)
\end{aligned}$$

are sharp given $\mathbf{n}_{C_1}$ *and* $\mathbf{n}_{C_2}$.

If the two marginals are overlapping, Proposition 5.1 says that the Fréchet bounds for $n_{D_0}(i^0_{D_0})$ are given by

$$\min\left\{n_{C_1\cap D_0}(i^0_{C_1\cap D_0}), n_{C_2\cap D_0}(i^0_{C_2\cap D_0})\right\}, \tag{15}$$

and

$$\max\left\{0, n_{C_1\cap D_0}(i^0_{C_1\cap D_0}) + n_{C_2\cap D_0}(i^0_{C_2\cap D_0}) - n_{C_1\cap C_2\cap D_0}(i^0_{C_1\cap C_2\cap D_0})\right\}. \tag{16}$$

It turns out that the bounds in equations (15) and (16) are *not* necessarily sharp bounds for $n_{D_0}(i^0_{D_0})$ given $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$.

Lemma 5.3 *Let the two fixed marginals $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ be such that $C_1 \cup C_2 = K$. Consider $D_0 \subset K$ and denote $D_1 := (C_1 \setminus C_2) \cap D_0$, $D_2 := (C_2 \setminus C_1) \cap D_0$ and $D_{12} := (C_1 \cap C_2) \cap D_0$. In addition, we let $C_{12} := (C_1 \cap C_2) \setminus D_0$. Then an upper bound for $n_{D_0}(i^0_{D_0})$ given $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ is:*

$$\sum_{i^1_{C_{12}} \in \mathcal{I}_{C_{12}}} \min\left\{n_{(C_1\cap D_0)\cup C_{12}}\left(i^0_{C_1\cap D_0}, i^1_{C_{12}}\right), n_{(C_2\cap D_0)\cup C_{12}}\left(i^0_{C_2\cap D_0}, i^1_{C_{12}}\right)\right\}, \tag{17}$$

while a lower bound is

$$\begin{aligned}\sum_{i^1_{C_{12}} \in \mathcal{I}_{C_{12}}} &\max\left\{0, n_{(C_1\cap D_0)\cup C_{12}}\left(i^0_{C_1\cap D_0}, i^1_{C_{12}}\right)\right. \\ &\left. + n_{(C_2\cap D_0)\cup C_{12}}\left(i^0_{C_2\cap D_0}, i^1_{C_{12}}\right) - n_{D_{12}}\left(i^0_{D_{12}}\right)\right\}.\end{aligned} \tag{18}$$

The following result characterizes the behavior of GSA in the decomposable case.

Proposition 5.2 *Let $\mathbf{n}$ be a k-dimensional table and consider the set of cells $\mathbf{T} = \mathbf{T}^{(\mathbf{n})}$ associated with $\mathbf{n}$ as it was defined in equation (4). The marginals $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$ induce a decomposable independence graph $\mathcal{G} = (K, E)$ with $\mathcal{C}(\mathcal{G}) = \{C_1, C_2, \ldots, C_p\}$ and $\mathcal{S}(\mathcal{G}) = \{S_2, \ldots, S_p\}$. The set of fixed cells $\mathbf{T}_0 \subset \mathbf{T}^{(\mathbf{n})}$ is given by the cell entries contained in the tables*

$$\bigcup_{r=1}^{p} \bigcup_{\{C: C\subseteq C_r\}} \mathcal{RD}(\mathbf{n}_C).$$

For every cell $t \in \mathbf{T}$, we let $\mathbf{n}_1^{(t)}$, $\mathbf{n}_2^{(t)}$, ..., $\mathbf{n}_{k_t}^{(t)}$ be the tables in $\mathcal{RD}$ such that t is a cell entry in $\mathbf{n}_r^{(t)}$, $r = 1, 2, \ldots, k_t$. Under these conditions, GSA converges to an upper bound $U_s(t)$ and to a lower bound $L_s(t)$ such that

$$\begin{aligned} &\max\{L^r(t) : r = 1, 2, \ldots, k_t\} \leq L_s(t), \\ &U_s(t) \leq \min\{U^r(t) : r = 1, 2, \ldots, k_t\}, \end{aligned} \tag{19}$$

where $U^r(t)$ and $L^r(t)$ are the Fréchet bounds of the cell t in table $\mathbf{n}_r^{(t)}$.

Any cell $t_0 \in \mathbf{T}$ can be found in one, two or possibly more tables in $\mathcal{RD}$. It is sufficient to prove that GSA converges to the Fréchet bounds for t_0 in every table $\mathbf{n}'$ such that t_0 is a cell of $\mathbf{n}'$. The shuttle procedure updates the bounds for t_0 once a better upper or lower bound is identified, so equation (19) is true if and only if the algorithm reaches the Fréchet bounds in every cell of every table in $\mathcal{RD}$. A cell $n(i^0)$, $i^0 \in \mathcal{I}$, might appear in several tables in $\mathcal{RD}$, but Proposition 5.2 implies that GSA converges to the Fréchet bounds in equation (12) of $n(i^0)$, and since from Theorem 5.1 we learn that these bounds are sharp, we deduce that the shuttle procedure reaches the sharp bounds for $n(i^0)$.

5.1 Example: Bounds for the Czech Autoworkers data (revisited)

We come back to the 2^6 contigency table given in Table 1. Whittaker (1990) suggests on page 263 that an appropriate model for this data is given by the marginals [BF], [ABCE] and [ADE]. This represents a decomposable log-linear model whose independence graph has separators [B] and [AE]. The corresponding Fréchet bounds from equation (12) become:

$$\begin{aligned} &\min\left\{n_{BF}\left(i_{BF}\right), n_{ABCE}\left(i_{ABCE}\right), n_{ADE}\left(i_{ADE}\right)\right\} \geq n(i) \geq \\ &\max\left\{n_{BF}\left(i_{BF}\right) + n_{ABCE}\left(i_{ABCE}\right) + n_{ADE}\left(i_{ADE}\right)\right. \\ &\left. -n_B\left(i_B\right) - n_{AE}\left(i_{AE}\right), 0\right\}. \end{aligned}$$

The bounds computed by GSA are shown in Table 3.

6 Computing sharp bounds

When the fixed set of marginals defines a decomposable independence graph, GSA converges to the corresponding Fréchet bounds for all the cell entries in the table $\mathbf{n}$. When $\mathbf{n}$ is dichotomous and all the lower-dimensional marginals are fixed, we were also able to explicitly determine the tightest bounds for the cells entries $\mathbf{n}$ and prove that GSA reaches these bounds. Even in these two particular instances GSA is guaranteed to find sharp bounds

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,246]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,151]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
pos	< 3	< 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]

Table 3: Bounds for entries in Table 1 induced by fixing the marginals [BF], [ABCE] and [ADE].

only for the cells $\mathcal{N}$ in table $\mathbf{n}$. In this section we present a method that sequentially adjusts the bounds $L_s(\mathbf{T})$ and $U_s(\mathbf{T})$ obtained from GSA until they become sharp.

The integer value $U(t_1)$ is a sharp upper bound for a cell $t_1 \in \mathbf{T}$ if and only if there exists an integer array $V(\mathbf{T}) \in S[L_s(\mathbf{T}), U_s(\mathbf{T})]$ with a count of $U(t_1)$ in cell t_1 (i.e., $V(t_1) = U(t_1)$) and if there does not exist another integer array $V'(\mathbf{T}) \in S[L_s(\mathbf{T}), U_s(\mathbf{T})]$ having a count in cell t_1 strictly bigger than $U(t_1)$ (i.e., $V'(t_1) > U(t_1)$). The sharp lower bound $L(t_1)$ can be defined in a similar way.

We know that $L_s(t_1) \leq L(t_1) \leq U(t_1) \leq U_s(t_1)$. This means that the first candidate value for $U(t_1)$ is $U_s(t_1)$. If there is no integer array $V(\mathbf{T}) \in S[L_s(\mathbf{T}), U_s(\mathbf{T})]$ with $V(t_1) = U_s(t_1)$, we sequentially try $U_s(t_1) - 1, U_s(t_1) - 2, \ldots, L_s(t_1)$ and stop when a feasible array with the corresponding count in cell t_1 is determined. The candidate values for the sharp lower bound $L(t_1)$ are $L_s(t_1) + 1, L_s(t_1) + 2, \ldots, U_s(t_1)$ in this particular order. After fixing the count $V(t_1)$ to an integer value between $L_s(t_1)$ and $U_s(t_1)$, we employ GSA to update the upper and lower bounds for all the cells in $\mathbf{T}$. Denote by $L_s^1(\mathbf{T})$ and $U_s^1(\mathbf{T})$ the new bounds identified by GSA. These bounds are tighter than $L_s(\mathbf{T})$ and $U_s(\mathbf{T})$, thus the set of integer arrays $S^1 = S[L_s^1(\mathbf{T}), U_s^1(\mathbf{T})]$ is included in $S[L_s(\mathbf{T}), U_s(\mathbf{T})]$. We reduced the problem of determining sharp bounds for the cell t_1 to the problem of checking whether S^1 is empty. We need to repeat these steps for every cell t_1 for which we want to obtain sharp bounds.

For notational simplicity we describe an algorithm for exhaustively enumerating all the

integer arrays in $S[L(\mathbf{T}), U(\mathbf{T})]$. Here $L(\mathbf{T})$ and $U(\mathbf{T})$ are arrays of lower and upper bounds for the cells $\mathbf{T}$. We associate with every cell $t = t_{J_1 J_2 \ldots J_k} \in \mathbf{T}$ an index (see Knuth (1973))

$$IND(t) := \sum_{l=1}^{k} 2^{\sum_{s=l+1}^{k} I_s} \left(\sum_{j_l \in J_l} 2^{j_l - 1} - 1 \right) + 1 \in \{1, 2, \ldots, N\},$$

and order the cells in $\mathbf{T}$ as a linear list $t_1, t_2, \ldots, t_N$, where $N = 2^{I_1 + I_2 + \ldots + I_k}$. With this ordering, we sequentially attempt to fix every cell at integer values between its current upper and lower bounds and use GSA to update the bounds for the remaining cells. We successfully determined a feasible array when we assigned a value to every cell and GSA did not identify any inconsistencies among these values.

PROCEDURE SharpBounds(k,$L^k(\mathbf{T})$,$U^k(\mathbf{T})$)

(1) IF $k = N+1$ THEN save the newly identified array $V(\mathbf{T}) \in S[L(\mathbf{T}), U(\mathbf{T})]$.

(2) FOR every integer $c \in [L^k(t_k), L^k(t_k)+1, \ldots, U^k(t_k)]$ DO

(2A) SET $V(t_k)$ to value c.

(2B) SET $L^{k+1}(t_k) = U^{k+1}(t_k) = c$, $L^{k+1}(t_i) = L^k(t_i)$, $U^{k+1}(t_i) = U^k(t_i)$ for $i = 1, \ldots, k-1, k+1, \ldots, N$.

(2C) Run GSA to update the bounds $L^{k+1}(\mathbf{T})$ and $U^{k+1}(\mathbf{T})$.

(2D) IF GSA did not identify any inconsistencies THEN
CALL SharpBounds(k+1,$L^{k+1}(\mathbf{T})$,$U^{k+1}(\mathbf{T})$).

PROCEDURE ENDS

The initial call is SharpBounds(1,$L(\mathbf{T})$,$U(\mathbf{T})$). Note that the updated bounds from step (2C) satisfy

$$L^k(t_i) \leq L^{k+1}(t_i) \leq U^{k+1}(t_i) \leq U^k(t_i),$$

provided that GSA did not report inconsistencies. This sequential improvement of the bounds avoids an exhaustive enumeration of all the combinations of possible values of the cells $\mathbf{T}$ that would lead to a very low computational efficiency of the algorithm.

When computing sharp bounds for a cell t_1, we can stop the SharpBounds procedure after we identified the first table in S^1 or learn that no such table exists.

7 Large Contingency Tables

We demonstrate the scalability of the generalized shuttle algorithm by computing sharp bounds for the non-zero entries of a a 2^{16} contingency table extracted from the "analytic" data file for National Long-Term Care Survey created by the Center of Demographic Studies at Duke University. Each dimension corresponds to a measure of disability defined by an activity of daily leaving, and the table contains information cross-classifying individuals aged 65 and above. The 16 dimensions of this contingency table correspond to six activities of daily living (ADLs) and ten instrumental activities of daily living (IADLs). Specifically, the ADLs are (1) eating, (2) getting in/out of bed, (3) getting around inside, (4) dressing, (5) bathing, (6) getting to the bathroom or using a toilet. The IADLs are (7) doing heavy house work, (8) doing light house work, (9) doing laundry, (10) cooking, (11) grocery shopping, (12) getting about outside, (13) travelling, (14) managing money, (15) taking medicine, (16) telephoning. For each ADL/IADL measure, subjects were classified as being either disabled (level 1) or healthy (level 0) on that measure. For a detailed description of this extract see Manton et al. (1993).

We applied GSA to compute sharp upper and lower bounds for the entries in this table corresponding to a number of different sets of fixed marginals. Here we describe one complex calculation for the set involving three fixed 15-way marginals obtained by collapsing the 16-way table across the variables (14) managing money, (15) taking medicine and (16) telephoning. Of the $2^{16} = 65,536$ cells, $62,384$ contain zero entries. Since the target table is so sparse, fixing three marginals of dimension fifteen leads to the exact determination (i.e., equal upper and lower bounds) of most of the cell entries. To be more exact, only 128 cells have the upper bounds strictly bigger than the lower bounds! The difference between the upper and lower bounds is equal to 1 for 96 cells, 2 for 16 cells, 6 for 8 cells, and 10 for 8 cells.

We take a closer look at the bounds associated with "small" counts of "1" or "2". There are $1,729$ cells containing a count of "1". Of these, $1,698$ cells have the upper bounds equal to the lower bounds. The difference between the bounds is 1 for 28 of the remaining counts of "1", is 2 for two other cells and is equal to 6 for only one entry. As for the 499 cells with a count of "2", the difference between the bounds is zero for 485 cells, is 1 for 10 cells and is 2 for 4 other cells.

GSA converged in approximately twenty iterations to the sharp bounds and it took less than six hours to complete on a single-processor machine at the Department of Statistics, Carnegie Mellon University. We re-checked these bounds by determining the feasible integer tables for which they are attained on the Terascale Computing System at the Pittsburgh

Supercomputing Center. We used a parallel implementation of GSA that independently adjust the bounds for various cells and the computations took almost one hour to complete on fifty-six processors.

8 Other examples

In the examples that follow we employ GSA not only to produce sharp bounds, but also to compute exact p-values for conditional inference with the hypergeometric distribution (see Dobra et al. (2006)):

$$p(\mathbf{n}) = \left[\prod_{i\in\mathcal{I}} n(i)!\right]^{-1} / \sum_{\mathbf{n}'\in\mathcal{T}} \left[\prod_{i\in\mathcal{I}} n'(i)!\right]^{-1}. \tag{20}$$

where $\mathcal{T}$ represents the set of contigency tables consistent with a given set of constraints (e.g., upper and lower bounds for cell entries). The corresponding p-value of the exact test is (see Guo and Thompson (1992)):

$$\sum_{\{\mathbf{n}'\in\mathcal{T}:p(\mathbf{n}')\leq p(\mathbf{n})\}} p(\mathbf{n}'), \tag{21}$$

where $\mathbf{n}$ is the observed table. Sundberg (1975) shows that the normalizing constant in equation (20) can be directly evaluated if $\mathcal{T}$ is determined by a decomposable set of marginals, but otherwise it can be computed only if $\mathcal{T}$ can be exhaustively enumerated. GSA can accomplish this task for almost any type of constraints and evaluate $p(\mathbf{n})$ as well as the p-value in equation (21) exactly.

We compare our inferences with the results obtained by Chen et al. (2006) who proposed a sequential importance sampling method (SIS, henceforth) for approximating exact p-values by randomly sampling from $\mathcal{T}$ and $p(\mathbf{n})$.

Example 1. Vlach (1986) considers the following three matrices:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Matrices A, B and C appear to be the 2-way marginals of a $6 \times 4 \times 3$ contingency table and their one-way marginals coincide; however, there does not exist a $6 \times 4 \times 3$ *integer* table having this set of two-way margins and GSA stopped without producing any bounds due to the inconsistencies it identified.

Example 2. Sullivant (2005) presented a $2\times2\times2\times2$ table with a grand total of 5, reproduced here as Table 4. This is the only integer table consistent with the six 2-way marginals and GSA correctly identifies it. Fitting the no 2nd-order interaction model implied by fixing the 2-way margins in R using `loglin` yields the MLEs in the right panel of Table 4, but the program reports $d.f. = 5$. The correct number of degrees of freedom is zero since there is only one integer table with these constraints. Testing the significance of the no second-order interaction model with reference to a χ^2 distribution on 5 degrees of freedom would be errorneous. The lower integer bounds equal the upper integer bounds for all 16 cells. Note the large gaps (up to 1.67) between the integer bounds and the real bounds (see Table 5) calculated with the simplex algorithm.

		C	No		Yes		No		Yes	
A	B	D	No	Yes	No	Yes	No	Yes	No	Yes
No	No		0	1	1	0	1.06	0.36	0.36	0.21
	Yes		1	0	0	0	0.36	0.21	0.21	0.21
Yes	No		1	0	0	0	0.36	0.21	0.21	0.21
	Yes		0	0	0	1	0.21	0.21	0.21	0.36

Table 4: A sparse four-way dichotomous table (left panel) from Sullivant (2005). The right panel gives the MLEs induced by the six 2-way marginals.

		C	No		Yes	
A	B	D	No	Yes	No	Yes
No	No		$[0, 1.67]$	$[0, 1]$	$[0, 1]$	$[0, 0.67]$
	Yes		$[0, 1]$	$[0, 0.67]$	$[0, 0.67]$	$[0, 0.67]$
Yes	No		$[0, 1]$	$[0, 0.67]$	$[0, 0.67]$	$[0, 0.67]$
	Yes		$[0, 0.67]$	$[0, 0.67]$	$[0, 0.67]$	$[0, 1]$

Table 5: LP bounds fixing the six 2-way marginals of Table 4.

Example 3. Dobra et al. (2006) used GSA to determine that there are 810 tables consistent with the set of fixed marginals [ACDEF], [ABDEF], [ABCDE], [BCDF], [ABCF], [BCEF] of Table 1. GSA calculates the p-value for the exact goodness-of-fit test in equation (21) to be 0.235. The estimated p-value computed using SIS in Chen et al. (2006) is 0.27, while the estimated number of tables is 840. By way of comparison, the `loglin` function in R gives a p-value of 0.21 on 4 degrees of freedom.

Dobra et al. (2006) also considered the model determined by fixing the 15 4-way margins. GSA reported a number of $705,884$ feasible tables with a correspinding exact p-value defined

by equation (21) equal to 0.432. Fitting the same model with `loglin` yields an approximate p-value of 0.438 by reference to a χ^2 distribution of 7.95 on 8 degrees of freedom.

			Age					
Race	Sex	Opinion	$18-25$	$26-35$	$36-45$	$46-55$	$56-65$	$66+$
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1
White	Male	Yes	$[90, 101]$	$[130, 146]$	$[107, 123]$	$[65, 81]$	$[61, 78]$	$[70, 87]$
		No	$[40, 49]$	$[58, 71]$	$[51, 63]$	$[43, 54]$	$[44, 57]$	$[55, 70]$
		Und	$[0, 2]$	$[0, 3]$	$[5, 9]$	$[4, 9]$	$[5, 9]$	$[7, 12]$
	Female	Yes	$[135, 146]$	$[163, 179]$	$[146, 162]$	$[95, 111]$	$[96, 113]$	$[107, 124]$
		No	$[38, 47]$	$[58, 71]$	$[51, 63]$	$[45, 56]$	$[50, 63]$	$[57, 72]$
		Und	$[0, 2]$	$[3, 6]$	$[6, 10]$	$[5, 10]$	$[7, 11]$	$[12, 17]$
Nonwhite	Male	Yes	$[19, 30]$	$[10, 26]$	$[10, 26]$	$[6, 22]$	$[0, 17]$	$[0, 17]$
		No	$[0, 9]$	$[0, 13]$	$[0, 12]$	$[0, 11]$	$[0, 13]$	$[0, 15]$
		Und	$[1, 3]$	$[0, 3]$	$[0, 4]$	$[0, 5]$	$[0, 4]$	$[0, 5]$
	Female	Yes	$[15, 26]$	$[17, 33]$	$[10, 26]$	$[7, 23]$	$[3, 10]$	$[0, 17]$
		No	$[0, 9]$	$[0, 13]$	$[0, 12]$	$[0, 11]$	$[0, 13]$	$[0, 15]$
		Und	$[0, 2]$	$[0, 3]$	$[0, 4]$	$[0, 5]$	$[0, 4]$	$[0, 5]$

Table 6: The upper panel gives the 4-way abortion option data from Haberman (1978). The lower panel gives the sharp integer bounds induced by the four 3-way marginals of this table.

Example 4. Table 6 contains a $2 \times 2 \times 3 \times 6$ table from an NORC survey from the 1970s (see Haberman (1978), page 291) that cross-classifies race (white, nonwhite), sex (male, female), attitude towards abortion (yes, no, undecided) and age ($18-25$, $26-35$, $36-45$, $46-55$, $56-65$, $66+$ years). Christensen (1997) page 111 considered the log-linear model corresponding to the four 3-way marginals. The `loglin` function in R yields an approximate p-value of 0.807 based on a χ^2 distribution on 6.09 on 10 degrees of freedom. GSA identified $83,087,976$ tables consistent with the 3-way marginals and returned an exact p-value for the goodness-of-fit test in equation (21) equal to 0.815. Chen et al. (2006) report that SIS estimated that the number of feasible tables is 9.1×10^7 and that the exact p-value based on the hypergeometric distribution is approximately 0.85. In the bottom panel of Table 6

C	S	R T	1	2	3
1	1	1	3	20	5
1	1	2	11	14	8
1	2	1	3	14	12
1	2	2	6	13	5
2	1	1	12	12	0
2	1	2	11	10	0
2	2	1	3	9	4
2	2	2	6	9	3

Table 7: Results of clinical trial for the effectiveness of an analgesic drug from Koch et al. (1983)

we give the upper and lower bounds computed by GSA. Note that the release of the four three-way marginals might be problematic from a disclosure limitation perspective due to the tight bounds for some of the small counts of 1 and 2.

Example 5. Table 7 from Koch et al. (1983) summarizes the results of a clinical trial on the effectiveness (R—poor, moderate or excellent) of an analgesic drug (T—1,2) for patients in two statuses (S) and two centers (C), center; S, status; T, treatment; R, response) with a grand total of 193. While most of the counts are relatively large, the table contains two counts of zero that lead to a zero entry in the [CSR] marginal.

Fienberg and Slavkovic (2004, 2005) discuss several log-linear models associated with this contingency table to illustrate disclosure limitation techniques. The upper and lower bounds presented in their 2004 paper are the same bounds identified by GSA, so we chose not to reproduce them again here. The zero entry in the [CSR] marginal leads to the non-existence of MLEs in any log-linear model with a generator [CSR]. This implies that the degrees of freedom for any log-linear model that includes [CSR] as a minimal sufficient statistic needs to be reduced by one — this corresponds to fitting a log-linear model to the incomplete table that does not include the two counts of zero adding up to the zero entry in the [CSR] marginal. For additional details and theoretical considerations, see Fienberg (1980) and Fienberg and Rinaldo (2007).

How does the exact goodness-of-fit test in equation (21) perform in this special situation? For the model [CST][CSR], GSA identifies $79,320,780$ feasible tables and gives an exact p-value of 0.073. By comparison, the `loglin` function in R yields an approximate p-value of 0.06 based on 7 degrees of freedom. For the model [CST][CSR][TR], GSA finds $155,745$ feasible tables with a corresponding p-value of 0.0499, while the `loglin` function gives a p-value of 0.039 based on 5 degrees of freedom. For the model [CST][CSR][CTR], GSA finds

$1,274$ feasible tables with a p-value of 0.152, while the `loglin` function reports a p-value of 0.127 based on 3 degrees of freedom. The last model we consider is [CST][CSR][SRT] with an exact p-value of 0.093 based on $1,022$ feasible tables. In this case `loglin` finds an approximate p-value of 0.073 based on 3 degrees of freedom. Remark that the discrepancy between the exact and approximate p-values tends to become more significant in degenerate cases when the MLEs do not exist. The model [CST][CSR][TR] seems to fit the data well indicating that there is evidence of a direct relationship between the treatment and response in this clinical trial.

Example 6. Dobra et al. (2008) analyze a sparse dichotomous 6-way table from Edwards (1992) which cross-classifies the parental alleles of six loci along a chromosome strand of a barley powder mildew fungus. The variables are labeled A, B, C, D, E and F, and have categories 1 or 2 – see Table 8. GSA finds a relatively small number $36,453$ of tables consistent with the two-way marginals with an exact p-value of the goodness-of-fit test based on the hypergeometric distribution equal to 0.652. The MLEs for this log-linear model do not exist because of a zero entry in the [AB] marginal; however, the MLEs for the log-linear model [ABCD][CDE][ABCEF] do exist. In this instance, GSA finds 30 tables consistent with the marginals [ABCD], [CDE] and [ABCEF] with an exact p-value of 1.

			D	1				2			
			E	1		2		1		2	
A	B	C	F	1	2	1	2	1	2	1	2
1	1	1		0	0	0	0	3	0	1	0
		2		0	1	0	0	0	1	0	0
	2	1		1	0	1	0	7	1	4	0
		2		0	0	0	2	1	3	0	11
2	1	1		16	1	4	0	1	0	0	0
		2		1	4	1	4	0	0	0	1
	2	1		0	0	0	0	0	0	0	0
		2		0	0	0	0	0	0	0	0
1	1	1		[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 4]	[0, 2]	[0, 1]	[0, 1]
		2		[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 2]	[0, 2]	[0, 1]	[0, 1]
	2	1		[0, 3]	[0, 2]	[0, 3]	[0, 2]	[0, 13]	[0, 2]	[0, 10]	[0, 2]
		2		[0, 1]	[0, 3]	[0, 2]	[0, 4]	[0, 2]	[0, 9]	[0, 2]	[2, 16]
2	1	1		[9, 22]	[0, 2]	[0, 9]	[0, 2]	[0, 2]	[0, 1]	[0, 2]	[0, 1]
		2		[0, 2]	[0, 10]	[0, 3]	[0, 10]	[0, 1]	[0, 2]	[0, 1]	[0, 2]
	2	1		[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]
		2		[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]	[0, 0]

Table 8: A sparse genetics 2^6 table from Edwards (1992). The upper panel gives the cell counts, while the lower panel shows to sharp bounds induced by fixing the two-way marginals.

9 Conclusions

We have described the Generalized Shuttle Algorithm (GSA) that exploits the hierarchical structure of categorical data to compute sharp bounds and enumerate sets of multi-way tables. The constraints defining these sets can come in the form of fixed marginals, upper and lower bounds on blocks of cells or structural zeros. In the most general setting one can restrict the search scope to tables having certain combinations of counts in various cell configurations. GSA produces sharp bounds not only for cells in the multi-way table analyzed, but also for any cells that belong to tables obtained through collapsing categories or variables. We showed through several examples that GSA performs very well and leads to valuable results.

We also illustrated that GSA can compute bounds for high-dimensional contingency tables. We are not aware how such computations can be performed through LP or IP methods. No matter how efficient LP/IP might be in solving one optimization problem, calculating bounds for a 16-dimensional table would involve solving $2 \times 2^{16} = 131,072$ separate optimzation problems and this represents a huge computational undertake. Instead, GSA computes bounds very close to the sharp bounds in one quick step, then adjusts these bounds to the sharp bounds only for the cells whose value is not uniquely determined by the marginal constraints.

While it is possible to increase the computational efficiency of GSA by adjusting the bounds in parallel or by choosing candidate values for the cell counts starting from the middle of the current feasibility intervals (see Dobra (2001)), we do not make any particular claims about its computational efficiency. The current implementation of the algorithm can be slow for a larger number of dimensions and categories and might need a lot of computer memory. On the other hand, GSA can easily be used as an off-the-shelf method for analyzing contingency tables since it is extremely flexible and does not require any additional input (e.g., Markov bases, LP bounds, etc) or intricate calibration heuristics. GSA is an excellent benchmark for judging the validity and performance of other related methods (e.g., SIS of Chen et al. (2006)) that have the potential to properly scale to high-dimensional data.

Acknowledgments

We thank Alessandro Rinaldo for his valuable comments. The preparation of this paper was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences, and NSF Grant DMS-0631589 and Army contract DAAD19-02-1-3-0389 to Carnegie Mellon University.

Appendix

Proof of Proposition 4.1

We start from the $(1,1,\ldots,1)$ cell and go through a sequence of cells $n(i)$ until we reach $n(i^0)$. We can write

$$\begin{aligned} n^* &= n_{C_l}(1,\ldots,1,i^0_{ql+1},\ldots,i^0_k) - n(1,\ldots,1,i^0_{ql},\ldots,i^0_k), \\ n(1,\ldots,1,i^0_{q_l},\ldots,i^0_k) &= n_{C_{(l-1)}}(1,\ldots,1,i^0_{q_{(l-1)}+1},\ldots,i^0_k) - n(1,\ldots,1,i^0_{q_{(l-1)}},\ldots,i^0_k), \\ &\vdots \\ n(1,\ldots,1,i^0_{q_2},\ldots,i^0_k) &= n_{C_1}(1,\ldots,1,i^0_{q_1+1},\ldots,i^0_k) - n(i^0). \end{aligned}$$

We add the above equalities to obtain equation (9).

Proof of Proposition 4.2

We write the equalities in the proof of Proposition 4.1 as

$$t_{\{1\}\ldots\{1\}} \oplus t_{\{1\}\ldots\{1\}\{i^0_{q_l}\}\ldots\{i^0_k\}} = t_{\{1\}\ldots\{i^0_{q_l-1}\}\{1,2\}\{i^0_{q_l+1}\}\ldots\{1\}},$$

and

$$t_{\{1\}\ldots\{1\}\{i^0_{q_{(s+1)}}\}\ldots\{i^0_k\}} \oplus t_{\{1\}\ldots\{1\}\{i^0_{q_s}\}\ldots\{i^0_k\}} = t_{\{1\}\ldots\{1\}\{i^0_{q_s-1}\}\{1,2\}\{i^0_{q_s+1}\}\ldots\{i^0_k\}},$$

for $s = 1,2,\ldots,l-1$. Hence

$$\left(t_{\{1\}\ldots\{1\}}, t_{\{1\}\ldots\{i^0_{q_l-1}\}\{1,2\}\{i^0_{q_l+1}\}\ldots\{1\}}, t_{\{1\}\ldots\{1\}\{i^0_{q_l}\}\ldots\{i^0_k\}}\right) \in \mathcal{Q}(\mathbf{T}),$$

and

$$\left(t_{\{1\}\ldots\{1\}\{i^0_{q_{(s+1)}}\}\ldots\{i^0_k\}}, t_{\{1\}\ldots\{1\}\{i^0_{q_s-1}\}\{1,2\}\{i^0_{q_s+1}\}\ldots\{i^0_k\}}, t_{\{1\}\ldots\{1\}\{i^0_{q_s}\}\ldots\{i^0_k\}}\right) \in \mathcal{Q}(\mathbf{T}),$$

for $s = 1,2,\ldots,l-1$. Since

$$\mathbf{T}'_0 := \left\{t_{\{1\}\ldots\{1\}\{i^0_{q_s-1}\}\{1,2\}\{i^0_{q_s+1}\}\ldots\{i^0_k\}} : s = 1,2,\ldots,l\right\} \subset \mathbf{T}_0,$$

the cells in $\mathbf{T}'_0$ have a fixed value

$$V\left(t_{\{1\}\ldots\{1\}\{i^0_{q_s-1}\}\{1,2\}\{i^0_{q_s+1}\}\ldots\{i^0_k\}}\right) n_{C_s}(1,\ldots,1,i^0_{q_s-1},i^0_{q_s+1},\ldots,i^0_k),$$

for $s = 1,2,\ldots,l$. GSA sequentially updates the bounds for the cells in $\mathbf{T}'_0$ in the following way:

$$\begin{aligned} L\left(t_{\{1\}\ldots\{1\}}\right) &= \max\left\{0, n_{C_l}(1,\ldots,1) - U\left(t_{\{1\}\ldots\{1\}\{i^0_{q_l}\}\ldots\{i^0_k\}}\right)\right\}, \\ U\left(t_{\ldots\{1\}\{i^0_{q_l}\}\ldots}\right) &= \min\left\{n_\emptyset, n_{C_{(l-1)}}(\ldots,1,i^0_{q_{(l-1)}+1},\ldots) - L\left(t_{\ldots\{1\}\{i^0_{q_{(l-1)}}\}\ldots}\right)\right\}, \\ &\vdots \end{aligned}$$

We set the non-negativity constraints

$$L\left(t_{\{1\}\ldots\{1\}\{i^0_{q_s}\}\ldots\{i^0_k\}}\right) \geq 0, \text{ for } s = 1, 2, \ldots, l, \tag{22}$$

then combine the above equalities to obtain equation (10). In an analogous manner we obtain the upper bounds in equation (11) from the identities:

$$\begin{aligned} U\left(t_{\{1\}\{1\}\ldots\{1\}}\right) &= \min\left\{n_\emptyset, n_{C_l}(1,\ldots,1) - L\left(t_{\{1\}\ldots\{1\}\{i^0_{q_l}\}\ldots\{i^0_k\}}\right)\right\}, \\ L\left(t_{\ldots\{1\}\{i^0_{q_l}\}\ldots}\right) &= \max\left\{0, n_{C_{(l-1)}}(\ldots,1,i^0_{q_{(l-1)}+1},\ldots) - U\left(t_{\ldots\{1\}\{i^0_{q_{(l-1)}}\}\ldots}\right)\right\}, \\ &\vdots \end{aligned}$$

Once GSA reaches the bounds in equations (10) and (11), no further changes are possible.

Proof of Proposition 5.1

The subgraph $\mathcal{G}(D)$ is decomposable since $\mathcal{G}$ is decomposable — see Lauritzen (1996). Equation (13) follows directly from Theorem 5.1 applied for table $\mathbf{n}_D$ which has a fixed set of marginals $\mathbf{n}_{C_1\cap D}$, $\mathbf{n}_{C_2\cap D}$, ..., $\mathbf{n}_{C_p\cap D}$. We clearly have $\mathcal{C}(\mathcal{G}(D)) = \{C_1\cap D, C_2\cap D, \ldots, C_p\cap D\}$ and $\mathcal{S}(\mathcal{G}(D)) = \{S_2\cap D, \ldots, S_p\cap D\}$.

Proof of Lemma 5.1

If $\mathcal{G}$ is complete, i.e. $p = 1$, we have $D_0 \subset K = C_1$, hence every entry $n_{D_0}(i^0_{D_0})$ will be fixed. Otherwise, it is known that $(\{v\}, \mathrm{bd}(v), V\setminus \mathrm{cl}(v))$ is a proper decomposition of $\mathcal{G}$ (Lauritzen, 1996). Since $\mathrm{bd}(v)$ is a separator of $\mathcal{G}$, X_v is independent of $X_{V\setminus\mathrm{cl}(v)}$ given $X_{\mathrm{bd}(v)}$. Therefore no information is lost if we think about $\mathbf{n}_{D_0}$ as being the marginal of $\mathbf{n}_{V\setminus\{v\}}$. The table $\mathbf{n}_{V\setminus\{v\}}$ has fixed marginals $\mathbf{n}_{C_1\setminus\{v\}}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$.

Proof of Lemma 5.2

The induced independence graph is obviously decomposable, and its cliques C_1 and C_2 are separated by the empty set. Every vertex $v \in (C_1\setminus D_0)\cup(C_2\setminus D_0)$ is simplicial in $\mathcal{G}$, hence we could think about $\mathbf{n}_{D_0}$ as being a table with two fixed non-overlapping marginals $\mathbf{n}_{C_1\cap D_0}$ and $\mathbf{n}_{C_2\cap D_0}$. Lemma 5.1 implies that we do not loose any information about the cell entry $n_{D_0}(i^0_{D_0})$ when collapsing across the variables $\{X_v : v \in (C_1\setminus D_0)\cup(C_2\setminus D_0)\}$. Thus the bounds in equation (14) are indeed sharp.

Proof of Lemma 5.3

If $C_{12} = \emptyset$, the bounds in equations (17) and (18) reduce to the Fréchet bounds in equations (15) and (16). We assume that $C_{12} \neq \emptyset$. The vertices in $C_1 \setminus (C_2 \cup D_0)$ and $C_2 \setminus (C_1 \cup D_0)$ are simplicial in the independence graph $\mathcal{G} = (K, E)$ induced by $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$. From Lemma 5.1, we deduce that we can restrict our attention to the marginal $\mathbf{n}_{D_0 \cup C_{12}}$ that has two fixed marginals $\mathbf{n}_{D_1 \cup (C_1 \cap C_2)} = \mathbf{n}_{(C_1 \cap D_0) \cup C_{12}}$ and $\mathbf{n}_{D_2 \cup (C_1 \cap C_2)} = \mathbf{n}_{(C_2 \cap D_0) \cup C_{12}}$. We choose an arbitrary index $i^1_{C_{12}} \in \mathcal{I}_{C_{12}}$. Consider the hyperplane $\mathbf{n}_{D_0}^{i^1_{C_{12}}}$ of $\mathbf{n}_{D_1 \cup (C_1 \cap C_2)}$ with entries

$$\mathbf{n}_{D_0}^{i^1_{C_{12}}}(i_{D_0}) := n_{D_0 \cup C_{12}}(i_{D_0}, i^1_{C_{12}}), \text{ for } i_{D_0} \in \mathcal{I}_{D_0}.$$

This hyperplane has two fixed marginals

$$\mathbf{n}_{C_1 \cap D_0}^{i^1_{C_{12}}} = \left\{ n_{(C_1 \cap D_0) \cup C_{12}} \left(i_{C_1 \cap D_0}, i^1_{C_{12}} \right) : i_{C_1 \cap D_0} \in \mathcal{I}_{C_1 \cap D_0} \right\},$$

and

$$\mathbf{n}_{C_2 \cap D_0}^{i^1_{C_{12}}} = \left\{ n_{(C_2 \cap D_0) \cup C_{12}} \left(i_{C_2 \cap D_0}, i^1_{C_{12}} \right) : i_{C_2 \cap D_0} \in \mathcal{I}_{C_2 \cap D_0} \right\}.$$

We have $D_0 = D_1 \cup D_{12} \cup D_2$, hence it is possible to make use of Theorem 5.1 to obtain the Fréchet bounds for the cell entry $n_{D_0}^{i^1_{C_{12}}}(i^0_{D_0}) = n_{D_0 \cup C_{12}}(i^0_{D_0}, i^1_{C_{12}})$, i.e.

$$\min \left\{ n_{(C_1 \cap D_0) \cup C_{12}} \left(i^0_{C_1 \cap D_0}, i^1_{C_{12}} \right), n_{(C_2 \cap D_0) \cup C_{12}} \left(i^0_{C_2 \cap D_0}, i^1_{C_{12}} \right) \right\},$$

and

$$\begin{aligned} \max \Big\{ 0, n_{(C_1 \cap D_0) \cup C_{12}} \left(i^0_{C_1 \cap D_0}, i^1_{C_{12}} \right) + \\ n_{(C_2 \cap D_0) \cup C_{12}} \left(i^0_{C_2 \cap D_0}, i^1_{C_{12}} \right) - n_{D_{12}} \left(i^0_{D_0 \cap D_{12}} \right) \Big\}. \end{aligned} \tag{23}$$

Since

$$n_{D_0}(i^0_{D_0}) = \sum_{i^1_{C_{12}} \in \mathcal{I}_{C_{12}}} n_{D_0 \cup C_{12}}(i^0_{D_0}, i^1_{C_{12}}),$$

Equations (17) and (18) follow from equation (23) by adding over all the indices $i^1_{C_{12}} \in \mathcal{I}_{C_{12}}$. Although the bounds in every hyperplane $\mathbf{n}_D^{i^1_{C_{12}}}$ are sharp, the bounds in equations (17) and (18) are guaranteed to be sharp only if $C_{12} = \emptyset$. If $C_{12} \neq \emptyset$, there is no reason to believe that equations (17) and (18) give sharp bounds for $\mathbf{n}_{D_0}(i^0_{D_0})$. Nevertheless, it is not hard to see that the upper bound in equation (15) is greater or equal to the upper bound in equation (17), while the lower bound in equation (16) is less or equal to the lower bound in equation (18). We conclude that the Fréchet upper and lower bounds for $\mathbf{n}_{D_0}(i^0_{D_0})$ are not necessarily the best bounds possible if $C_{12} \neq \emptyset$.

Proof of Proposition 5.2

We prove Proposition 5.2 by sequentially considering several particular cases. First we show that the shuttle procedure obtains the Fréchet bounds for a 2×2 table. Since any two-way table can be reduced to a number of 2×2 tables, it follows that the Fréchet bounds are also attained for a two-dimensional cross-classification with fixed one-dimensional totals. By induction on the number of fixed marginals of an arbitrary k-dimensional table $\mathbf{n}$ and by exploiting the fact that, if $\mathbf{n}$ has two marginals fixed, $\mathbf{n}$ can be split in several two-way tables with fixed one-way marginals, we are able to prove in the last subsection that the Fréchet bounds are attained for decomposable log-linear models with any number of minimal sufficient statistics.

The 2×2 case

Consider a 2×2 table $\mathbf{n} = \{n_{ij} : 1 \leq i, j \leq 2\}$ with fixed row totals $\{n_{1+}, n_{2+}\}$ and column totals $\{n_{+1}, n_{+2}\}$. The grand total of the table is n_{++}. The set $\mathbf{T}$ associated with $\mathbf{n}$ is given by

$$\mathbf{T} \quad = \quad \{n_{11}, n_{12}, n_{21}, n_{22}, n_{1+}, n_{2+}, n_{+1}, n_{+2}, n_{++}\}, \tag{24}$$

while the set of cells having a fixed value is $\mathbf{T}_0 = \{n_{1+}, n_{2+}, n_{+1}, n_{+2}, n_{++}\}$. There are only six dependencies

$$\begin{aligned} \mathcal{Q}(\mathbf{T})\{&(n_{1+}, n_{++}, n_{2+}), (n_{+1}, n_{++}, n_{+2}), (n_{11}, n_{1+}, n_{12}), \\ &(n_{12}, n_{+2}, n_{22}), (n_{21}, n_{2+}, n_{22}), (n_{11}, n_{21}, n_{+1})\}. \end{aligned} \tag{25}$$

The first two dependencies are redundant because they involve only the cells in $\mathbf{T}_0$. We show that GSA converges to the Fréchet bounds:

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}, \text{ for } 1 \leq i, j \leq 2. \tag{26}$$

We initialize the upper and lower bounds of the four cells in $\mathbf{T} \setminus \mathbf{T}_0$:

$$\begin{aligned} L(n_{11}) = L(n_{12}) = L(n_{21}) = L(n_{22}) \quad &:= \quad 0, \text{ and} \\ U(n_{11}) = U(n_{12}) = U(n_{21}) = U(n_{22}) \quad &:= \quad n_{++}. \end{aligned} \tag{27}$$

We sequentially go through the dependencies in $\mathcal{Q}(\mathbf{T})$. When we obtain a Fréchet bound, we mark it with "$\diamond$". Since the Fréchet bounds are sharp, once GSA reaches such a bound, it stays at that bound.

First iteration, dependency: $n_{11} \oplus n_{12} = n_{1+}$.

$$\begin{aligned}
L(n_{11}) &= \max\{L(n_{11}), n_{1+} - U(n_{12})\} = \max\{0, n_{1+} - n_{++}\} = 0, \\
U(n_{11}) &= \min\{U(n_{11}), n_{1+} - L(n_{12})\} = \min\{n_{++}, n_{1+}\} = n_{1+}, \\
L(n_{12}) &= \max\{L(n_{12}), n_{1+} - U(n_{11})\} = \max\{0, n_{1+} - n_{1+}\} = 0. \\
U(n_{12}) &= \min\{U(n_{12}), n_{1+} - L(n_{11})\} = \min\{n_{++}, n_{1+}\} = n_{1+}.
\end{aligned}$$

First iteration, dependency: $n_{12} \oplus n_{22} = n_{+2}$.

$$\begin{aligned}
L(n_{22}) &= \max\{L(n_{22}), n_{+2} - U(n_{12})\}, \\
&= \max\{0, n_{+2} - n_{1+}\} = \max\{0, n_{+2} + n_{2+} - n_{++}\}.\diamond \\
U(n_{22}) &= \min\{U(n_{22}), n_{+2} - L(n_{12})\}, \\
&= \min\{n_{++}, n_{+2}\} = n_{+2}. \\
L(n_{12}) &= \max\{L(n_{12}), n_{+2} - U(n_{22})\} = 0. \\
U(n_{12}) &= \min\{U(n_{12}), n_{+2} - L(n_{22})\}, \\
&= \min\{n_{1+}, n_{+2} + \min\{0, n_{1+} - n_{+2}\}\} = \min\{n_{+2}, n_{1+}\}.\diamond
\end{aligned}$$

First iteration, dependency: $n_{21} \oplus n_{22} = n_{2+}$.

$$\begin{aligned}
L(n_{21}) &= \max\{L(n_{21}), n_{2+} - U(n_{22})\}, \\
&= \max\{0, n_{2+} - n_{+2}\} = \max\{0, n_{2+} + n_{+1} - n_{++}\}.\diamond \\
U(n_{21}) &= \min\{U(n_{21}), n_{2+} - L(n_{22})\}, \\
&= \min\{0, n_{2+} + \min\{0, n_{1+} - n_{+2}\}\}, \\
&= \min\{n_{2+}, n_{++} - n_{+2}\} = \min\{n_{2+}, n_{+1}\}.\diamond \\
U(n_{22}) &= \min\{U(n_{22}), n_{2+} - L(n_{21})\}, \\
&= \min\{n_{+2}, n_{2+} + \min\{0, n_{+2} - n_{2+}\}\} = \min\{n_{+2}, n_{2+}\}.\diamond
\end{aligned}$$

First iteration, dependency: $n_{11} \oplus n_{21} = n_{+1}$.

$$\begin{aligned}
L(n_{11}) &= \max\{L(n_{11}), n_{+1} - U(n_{21})\}, \\
&= \max\{0, n_{+1} + \max\{-n_{2+}, -n_{+1}\}\}, \\
&= \max\{0, n_{+1} - n_{2+}\} = \max\{0, n_{+1} + n_{1+} - n_{++}\}.\diamond \\
U(n_{11}) &= \min\{U(n_{11}), n_{+1} - L(n_{21})\}, \\
&= \min\{n_{1+}, n_{+1} + \max\{-n_{2+}, -n_{+1}\}\}, \\
&= \min\{n_{+1}, n_{++} - n_{2+}\} = \min\{n_{+1}, n_{1+}\}.\diamond
\end{aligned}$$

Second iteration, dependency: $n_{11} \oplus n_{12} = n_{1+}$.

$$\begin{aligned} L(n_{12}) &= \max\{L(n_{12}), n_{1+} - U(n_{11})\}, \\ &= \max\{0, n_{1+} + \max\{-n_{+1}, -n_{1+}\}\}, \\ &= \max\{0, n_{1+} - n_{+1}\} = \max\{0, n_{1+} + n_{+2} - n_{++}\}.\diamond \end{aligned}$$

We see that the Fréchet bounds in equation (26) for all four cells in table $\mathbf{n}$ are obtained before completing the second iteration. Therefore Proposition 5.2 holds for any 2×2 table.

The two-way case

The next step is to examine a two-dimensional table $\mathbf{n} = \{n_{ij} : 1 \le i \le I_1, 1 \le j \le I_2\}$ for some $I_1, I_2 \ge 2$. This table has fixed row sums and column sums:

$$\{n_{i+} : 1 \le i \le I_1\} \cup \{n_{+j} : 1 \le j \le I_2\} \cup \{n_{++}\} \subset \mathbf{T}_0. \tag{28}$$

More precisely, the set of fixed cells $\mathbf{T}_0 \subset \mathbf{T} = \mathbf{T}^{(\mathbf{n})}$ is

$$\begin{aligned} \mathbf{T}_0 = \left\{t_{J_1\{1,2,\ldots,I_2\}} : \emptyset \ne J_1 \subseteq \{1, 2, \ldots, I_1\}\right\} \cup \\ \left\{t_{\{1,2,\ldots,I_1\}J_2} : \emptyset \ne J_2 \subseteq \{1, 2, \ldots, I_2\}\right\}. \end{aligned}$$

We remark that $t \in \mathbf{T} \setminus \mathbf{T}_0$ if and only if t is a cell in a table $\mathbf{n}' \in \mathcal{RD}(\mathbf{n})$. In other words, t does not have a fixed value if and only if t is a cell in a table of dimension 2 that could be obtained from $\mathbf{n}$ by table redesign. We show that the Fréchet bounds from Theorem 5.1 are attained when running GSA for every cell n_{ij}. It is sufficient to prove it for the $(1, 1)$ cell. The Fréchet inequality for n_{11} is

$$\min\{n_{1+}, n_{+1}\} \ge n_{11} \ge \max\{0, n_{1+} + n_{+1} - n_{++}\}. \tag{29}$$

We notice the similarity of equation (29) with equation (26). We define the 2×2 table $\mathbf{n}' = \{n'_{ij} : 1 \le i, j \le 2\}$, where

$$n'_{11} := n_{11},\ n'_{12} := \sum_{j>1} n_{1j},\ n'_{21} := \sum_{i>1} n_{i1},\ n'_{22} := \sum_{i>1}\sum_{j>1} n_{ij}. \tag{30}$$

This table has fixed row totals $\left\{n_{1+}, \sum_{i>1} n_{i+}\right\}$ as well as fixed column totals $\left\{n_{+1}, \sum_{j>1} n_{+j}\right\}$. The Fréchet bounds for the $(1, 1)$ count in table $\mathbf{n}'$ coincide with the Fréchet bounds for the $(1, 1)$ count in table $\mathbf{n}$. Since the four cells in table $\mathbf{n}'$ are also cells in the set $\mathbf{T}$ associated with $\mathbf{n}$, the generalized shuttle algorithm employed for the table $\mathbf{n}$ is equivalent to the shuttle procedure employed for the table $\mathbf{n}'$ from the perspective of finding sharp bounds for

$\{n'_{11}, n'_{12}, n'_{21}, n'_{22}\}$. We proved before that the generalized shuttle algorithm will converge to the Fréchet bounds for any 2×2 table, hence GSA finds the Fréchet bounds for the $(1,1)$ cell in table $\mathbf{n}$.

Now take an arbitrary cell $t = t_{\{i_1,i_2,\ldots,i_l\}\{j_1,j_2,\ldots,j_s\}} \in \mathbf{T} \setminus \mathbf{T}_0$. Consider the 2×2 table $\mathbf{n}^{(t)}$ with entries

$$\begin{aligned} \big\{ & t_{\{i_1,i_2,\ldots,i_l\}\{j_1,j_2,\ldots,j_s\}}, t_{\{i_1,i_2,\ldots,i_l\}(\mathcal{I}_2 \setminus \{j_1,j_2,\ldots,j_s\})}, \\ & t_{(\mathcal{I}_1 \setminus \{i_1,i_2,\ldots,i_l\})\{j_1,j_2,\ldots,j_s\}}, t_{(\mathcal{I}_1 \setminus \{i_1,i_2,\ldots,i_l\})(\mathcal{I}_2 \setminus \{j_1,j_2,\ldots,j_s\})} \big\}. \end{aligned}$$

The Fréchet bounds for the value $V(t)$ of cell t in the above table are

$$\min \left\{ V(t_{\{i_1,i_2,\ldots,i_l\}\{1,2,\ldots,I_2\}}), V(t_{\{1,2,\ldots,I_1\}\{j_1,j_2,\ldots,j_s\}}) \right\}$$

and

$$\begin{aligned} \max \big\{ 0, & V(t_{\{i_1,\ldots,i_l\}\{1,\ldots,I_2\}}) + V(t_{\{1,\ldots,I_1\}\{j_1,\ldots,j_s\}}) \\ & - V(t_{\{1,\ldots,I_1\}\{1,\ldots,I_2\}}) \big\}. \end{aligned} \tag{31}$$

The table $\mathbf{n}^{(t)}$ has fixed one-dimensional totals, hence we know the cell values

$$\begin{aligned} V(t_{\{i_1,i_2,\ldots,i_l\}\{1,2,\ldots,I_2\}}) &= \sum_{r=1}^{l} n_{i_r+}, \\ V(t_{\{1,2,\ldots,I_1\}\{j_1,j_2,\ldots,j_s\}}) &= \sum_{r=1}^{s} n_{+j_r}, \\ V(t_{\{1,2,\ldots,I_1\}\{1,2,\ldots,I_2\}}) &= n_{\emptyset}. \end{aligned}$$

The Fréchet bounds in equation (31) are the Fréchet bounds associated with cell t in every table $\mathbf{n}' \in \mathcal{RD}$ such that t is a cell in $\mathbf{n}'$. Again, for every such table $\mathbf{n}'$, it is true that $\mathbf{T}^{(\mathbf{n}')} \subset \mathbf{T}^{(\mathbf{n})}$ and $\mathcal{Q}(\mathbf{T}^{(\mathbf{n}')}) \subset \mathcal{Q}(\mathbf{T}^{(\mathbf{n})})$. When employing the shuttle procedure for $\mathbf{n}$ we also run the shuttle procedure in $\mathbf{n}'$, thus the bounds in equation (31) are attained by GSA and hence Proposition 5.2 holds for an arbitrary two-dimensional table.

Bounds induced by two fixed marginals

Let $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$ be a k-dimensional frequency count table having fixed marginals $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ such that $C_1 \cup C_2 = K$. The Fréchet bounds for a cell entry $n(i^0)$ are

$$\begin{aligned} \min \left\{ n_{C_1}\left(i^0_{C_1}\right), n_{C_2}\left(i^0_{C_2}\right) \right\} &\geq n(i^0), \\ n(i^0) \geq \min \left\{ n_{C_1}\left(i^0_{C_1}\right) + n_{C_2}\left(i^0_{C_2}\right) - n_{C_1 \cap C_2}\left(i^0_{C_1 \cap C_2}\right) \right\}. & \end{aligned}$$

First we study the case when the fixed marginals are non-overlapping. i.e. $C_1 \cap C_2 = \emptyset$. We attempt to reduce this case to the case of two-dimensional tables we studied before for which we know that Proposition 5.2 is true. The above inequalities become

$$\begin{aligned} \min\left\{n_{C_1}\left(i^0_{C_1}\right), n_{K\setminus C_1}\left(i^0_{K\setminus C_1}\right)\right\} &\geq n(i^0), \\ n(i^0) &\geq \min\left\{n_{C_1}\left(i^0_{C_1}\right) + n_{K\setminus C_1}\left(i^0_{K\setminus C_1}\right) - n_\emptyset\right\}. \end{aligned} \tag{32}$$

Without restricting the generality, we can assume that $C_1 = \{1, \ldots, l\}$ and $C_2 = \{l+1, \ldots, k\}$. To every index $i_{C_1} = (i_1, \ldots, i_l) \in \mathcal{I}_{C_1}$ we define (see Knuth (1973)):

$$IND_{C_1}(i_{C_1}) := \sum_{r=1}^{l}\left[\prod_{s=r+1}^{l} I_s\right]\cdot(i_r - 1) + 1 \in \{1, \ldots, I_1 \cdot I_2 \cdot \ldots \cdot I_l\}.$$

IND_{C_1} induces a one-to-one correspondence between the sets $\mathcal{I}_{C_1}$ and $\{1, \ldots, I_1 \cdot \ldots \cdot I_l\}$. Similarly, to every $i_{C_2} = (i_{l+1}, \ldots, i_k) \in \mathcal{I}_{C_2}$, we assign

$$IND_{C_2}(i_{C_2}) := \sum_{r=l+1}^{k}\left[\prod_{s=r+1}^{k} I_s\right]\cdot(i_r - 1) + 1 \in \{1, \ldots, I_{l+1} \cdot \ldots \cdot I_k\}.$$

Introduce two new compound variables Y_1 and Y_2 that take values in the sets $\{1, \ldots, I_1 \cdot I_2 \cdot \ldots \cdot I_l\}$ and $\{1, \ldots, I_{l+1} \cdot \ldots \cdot I_k\}$, respectively. Consider a two-way table

$$\mathbf{n}' = \{n'_{j_1 j_2} : 1 \leq j_1 \leq I_1 \cdot I_2 \cdot \ldots \cdot I_l, 1 \leq j_2 \leq I_{l+1} \cdot \ldots \cdot I_k\}$$

with entries given by

$$n'_{j_1 j_2} = n_K\left(IND^{-1}_{C_1}(j_1), IND^{-1}_{C_2}(j_2)\right).$$

The table $\mathbf{n}'$ has fixed row totals

$$\{n_{j_1+} : 1 \leq j_1 \leq I_1 \cdot I_2 \cdot \ldots \cdot I_l\},$$

where $n_{j_1+} n_{C_1}\left(IND^{-1}_{C_1}(j_1)\right)$, and column totals

$$\{n_{+j_2} : 1 \leq j_2 \leq I_{l+1} \cdot \ldots \cdot I_k\},$$

where $n_{+j_2} n_{C_2}\left(IND^{-1}_{C_2}(j_2)\right)$. Therefore there is a one-to-one correspondence between the cells in the original k-dimensional table $\mathbf{n}$ and the cells in the two-way table $\mathbf{n}'$. Moreover, there is a one-to-one correspondence between the fixed cells in $\mathbf{n}$ and the set of fixed cells in $\mathbf{n}'$. Running GSA for $\mathbf{n}$ assuming fixed marginals $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ is the same as running the shuttle procedure for $\mathbf{n}'$ assuming fixed one-dimensional totals. This implies that the Fréchet bounds in equation (32) are attained.

Consider a cell $t \in \mathbf{T} \setminus \mathcal{N}$ and let $\mathbf{n}' \in \mathcal{RD}$ as it was defined in equation (2) such that $t = n'(i^0)$, for some $i^0 \in \mathcal{I}'_1 \times \mathcal{I}'_2 \times \ldots \times \mathcal{I}'_k$. If $\mathbf{n}' \in \mathcal{RD}(\mathbf{n})$, then the Fréchet bounds for $t = n'(i^0)$ in table $\mathbf{n}'$ are

$$\begin{aligned} \min\left\{n'_{C_1}\left(i^0_{C_1}\right), n'_{K\setminus C_1}\left(i^0_{K\setminus C_1}\right)\right\} &\geq n'(i^0), \\ n'(i^0) \geq \min\left\{n'_{C_1}\left(i^0_{C_1}\right) + n'_{K\setminus C_1}\left(i^0_{K\setminus C_1}\right) - n_\emptyset\right\}&. \end{aligned} \tag{33}$$

$\mathbf{n}'_{C_1}$ and $\mathbf{n}'_{K\setminus C_1}$ are fixed marginals of $\mathbf{n}'$ obtained from $\mathbf{n}_{C_1}$ and $\mathbf{n}_{K\setminus C_1}$ by the same sequence of "category-join" operations that was necessary to transform the initial table $\mathbf{n}$ in $\mathbf{n}'$. Again, we have $\mathbf{T}^{(\mathbf{n}')} \subset \mathbf{T}^{(\mathbf{n})}$ and $\mathcal{Q}(\mathbf{T}^{(\mathbf{n}')}) \subset \mathcal{Q}(\mathbf{T}^{(\mathbf{n})})$, thus the Fréchet bounds in equation (33) are obtained by employing the shuttle procedure for the same reasons the bounds in equation (32) were reached.

Now assume that $\mathbf{n}' = \mathbf{n}_{D_0}$, $D_0 \subset K$, with $t = n_{D_0}(i^0_{D_0})$ for some $i^0_{D_0} \in \mathcal{I}_{D_0}$. The Fréchet bounds in $\mathbf{n}_{D_0}$ are given in Lemma 5.2. The table $\mathbf{n}_{D_0}$ has two fixed non-overlapping marginals $\mathbf{n}_{C_1\cap D_0}$ and $\mathbf{n}_{C_2\cap D_0}$, hence GSA reaches the Fréchet bounds in equation (14) because
$\mathbf{T}^{(\mathbf{n}_{D_0})} \subset \mathbf{T}^{(\mathbf{n})}$ and $\mathcal{Q}(\mathbf{T}^{(\mathbf{n}_{D_0})}) \subset \mathcal{Q}(\mathbf{T}^{(\mathbf{n})})$. If $\mathbf{n}' \in \mathcal{RD}(\mathbf{n}_{D_0})$ $\mathbf{n}'$ has two fixed marginals $\mathbf{n}'_{C_1\cap D_0}$ and $\mathbf{n}'_{C_2\cap D_0}$ obtained from $\mathbf{n}_{C_1\cap D_0}$ and $\mathbf{n}_{C_2\cap D_0}$ by joining categories associated with the variables cross-classified in $\mathbf{n}$. It is sufficient to replace $\mathbf{n}_{D_0}$ with $\mathbf{n}'$ in equation (14) to calculate the Fréchet bounds for t in table $\mathbf{n}'$.

If the two fixed marginals are overlapping, we can assume that there exist q and l with $1 \leq q \leq l \leq k$, such that $C_1 = \{1, 2, \ldots, l\}$ and $C_2 = \{q, q+1, \ldots, k\}$. Then $C_1 \cap C_2 = \{q, \ldots, l\}$. We reduce the case of two fixed overlapping marginals to the case of two fixed non-overlapping marginals by decomposing the tables $\mathbf{n}$, $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ in a number of hyperplanes. Each hyperplane of $\mathbf{n}$ has two non-overlapping marginals that are hyperplanes of $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$. Denote

$$D_1 := C_1 \setminus C_2 = \{1, 2, \ldots, q-1\}, \text{ and } D_2 := C_2 \setminus C_1 = \{l+1, l+2, \ldots, k\}.$$

Take the set of contingency tables

$$\left\{\mathbf{n}^{i^0_q,\ldots,i^0_l} = \left\{n^{i^0_q,\ldots,i^0_l}\left(i_{D_1\cup D_2}\right) : i_{D_1\cup D_2} \in \mathcal{I}_{D_1\cup D_2}\right\} : i^0_q \in \mathcal{I}_q, \ldots, i^0_l \in \mathcal{I}_l\right\},$$

where

$$\begin{aligned} n^{i^0_q,\ldots,i^0_l}(i_{D_1\cup D_2}) &= n^{i^0_q,\ldots,i^0_l}(i_1, \ldots, i_{q-1}, i_{l+1}, \ldots, i_k) \\ &= n(i_1, \ldots, i_{q-1}, i^0_q, \ldots, i^0_l, i_{l+1}, \ldots, i_k). \end{aligned}$$

Every table $\mathbf{n}^{i^0_q,\ldots,i^0_l}$ has two fixed non-overlapping marginals

$$\mathbf{n}^{i^0_q,\ldots,i^0_l}_{D_1} = \left\{n^{i^0_q,\ldots,i^0_l}\left(i_{D_1}\right) : i_{D_1} \in \mathcal{I}_{D_1}\right\},$$

with entries given by

$$\begin{aligned} n^{i_q^0,\ldots,i_l^0}\left(i_{D_1}\right) &= n^{i_q^0,\ldots,i_l^0}\left(i_1,\ldots,i_{q-1}\right) \\ &= n_{C_1}\left(i_1,\ldots,i_{q-1},i_q^0,\ldots,i_l^0\right), \end{aligned}$$

and $\mathbf{n}_{D_2}^{i_q^0,\ldots,i_l^0} = \left\{ n^{i_q^0,\ldots,i_l^0}\left(i_{D_2}\right) : i_{D_2} \in \mathcal{I}_{D_2} \right\}$, with entries given by

$$n^{i_q^0,\ldots,i_l^0}\left(i_{D_2}\right) n^{i_q^0,\ldots,i_l^0}\left(i_{l+1},\ldots,i_k\right) = n_{C_2}\left(i_q^0,\ldots,i_l^0,i_{l+1},\ldots,i_k\right).$$

Notice that the table $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ is a hyperplane of the original table $\mathbf{n}$, whereas $\mathbf{n}_{D_1}^{i_q^0,\ldots,i_l^0}$ is a hyperplane of $\mathbf{n}_{C_1}$, and $\mathbf{n}_{D_2}^{i_q^0,\ldots,i_l^0}$ is a hyperplane of $\mathbf{n}_{C_2}$. Employing the generalized shuttle algorithm for $\mathbf{n}$ is equivalent to employing distinct versions of the shuttle procedure for every hyperplane determined by an index $\left(i_q^0,\ldots,i_l^0\right) \in \mathcal{I}_{C_1\cap C_2}$. We already showed that GSA for $\mathbf{n}^{i_q^0,\ldots,i_l^0}$ converges to the Fréchet bounds of the cell entry $n^{i_q^0,\ldots,i_l^0}\left(i_{D_1}^0,i_{D_2}^0\right)$ (compare with equation (32)):

$$\begin{aligned} &\min\left\{ n_{D_1}^{i_q^0,\ldots,i_l^0}\left(i_{D_1}^0\right), n_{D_2}^{i_q^0,\ldots,i_l^0}\left(i_{D_2}^0\right) \right\} \geq \\ n^{i_q^0,\ldots,i_l^0}\left(i_{D_1}^0,i_{D_2}^0\right) &\geq \max\left\{ n_{D_1}^{i_q^0,\ldots,i_l^0}\left(i_{D_1}^0\right) + n_{D_2}^{i_q^0,\ldots,i_l^0}\left(i_{D_2}^0\right) - n_{\emptyset}^{i_q^0,\ldots,i_l^0} \right\}, \end{aligned} \tag{34}$$

where $n_{\emptyset}^{i_q^0,\ldots,i_l^0} = n_{C_1\cap C_2}\left(i_q^0,\ldots,i_l^0\right)$ is the grand total of the hyperplane $\mathbf{n}^{i_q^0,\ldots,i_l^0}$. Equation (34) can equivalently be written as

$$\begin{aligned} &\min\left\{ n_{C_1}\left(i_{D_1}^0,i_q^0,\ldots,i_l^0\right), n_{C_2}\left(i_q^0,\ldots,i_l^0,i_{D_2}\right) \right\} \geq n\left(i_{D_1}^0,i_q^0,\ldots,i_l^0,i_{D_2}^0\right) \\ &\quad \geq \max\left\{ 0, n_{C_1}\left(i_{D_1}^0,i_{C_1\cap C_2}^0\right) + n_{C_2}\left(i_{C_1\cap C_2}^0,i_{D_2}^0\right) - n_{C_1\cap C_2}\left(i_{C_1\cap C_2}^0\right) \right\}. \end{aligned}$$

These inequalities represent the Fréchet bounds for the cell count

$$n\left(i^0\right) = n\left(i_{D_1}^0,i_q^0,\ldots,i_l^0,i_{D_2}^0\right).$$

Now we show that any table $\mathbf{n}' \in \mathcal{RD} \setminus \bigcup_{r=1}^{2} \bigcup_{\{C:C\subseteq C_r\}} \mathcal{RD}(\mathbf{n}_C)$ can be separated in a number of hyperplanes such that the two fixed marginals of every hyperplane are non-overlapping. Let $\mathbf{n}'$ as it was defined in equation (2) and consider an arbitrary cell in $\mathbf{n}'$ specified by the index $(J_1^0,\ldots,J_k^0) \in \mathcal{I}_1' \times \ldots \times \mathcal{I}_k'$. The hyperplane $\mathbf{n}'^{(J_q^0,\ldots,J_l^0)}$ of table $\mathbf{n}'$ has entries

$$\left\{ n'(J_1,\ldots,J_{q-1},J_q^0,\ldots,J_l^0,J_{l+1},\ldots,J_k) : J_r \in \mathcal{I}_r' \right\},$$

for $r = 1,\ldots,q-1,l+1,\ldots,k$. The fixed overlapping marginals $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ induce two fixed overlapping marginals $\mathbf{n}'_{C_1}$ and $\mathbf{n}'_{C_2}$ of $\mathbf{n}'$. The index set of $\mathbf{n}'_{C_r}$, $r = 1,2$, is $\mathcal{I}'_{1;C_r} \times \ldots \times \mathcal{I}'_{k;C_r}$, where

$$\mathcal{I}'_{s;C_r} = \begin{cases} \mathcal{I}'_s, & \text{if } s \in C_r, \\ \{\mathcal{I}_s\}, & \text{if } s \notin C_r. \end{cases}$$

We define the hyperplanes $\mathbf{n'}_{C_1}^{(J_q^0,\ldots,J_l^0)}$ of $\mathbf{n}'_{C_1}$ and $\mathbf{n'}_{C_2}^{(J_q^0,\ldots,J_l^0)}$ of $\mathbf{n}'_{C_2}$ in the same way we defined the hyperplane $\mathbf{n'}^{(J_q^0,\ldots,J_l^0)}$ of $\mathbf{n}'$. Therefore $\mathbf{n'}^{(J_q^0,\ldots,J_l^0)}$ is a table having two fixed non-overlapping marginals $\mathbf{n'}_{C_1}^{(J_q^0,\ldots,J_l^0)}$ and $\mathbf{n'}_{C_2}^{(J_q^0,\ldots,J_l^0)}$. The Fréchet bounds for $n'(J_1^0,\ldots,J_k^0)$ coincide with the Fréchet bounds for the cell entry

$$n'^{(J_q^0,\ldots,J_l^0)}(J_1^0,\ldots,J_{q-1}^0,J_{l+1}^0,\ldots,J_k^0)$$

in table $\mathbf{n'}^{(J_q^0,\ldots,J_l^0)}$. Therefore Proposition 5.2 holds for any table of counts with two fixed marginals.

Calculating bounds in the general decomposable case

The set of fixed cliques defines a decomposable independence graph $\mathcal{G}=(K,E)$ with cliques $\mathcal{C}(\mathcal{G})$ and separators $\mathcal{S}(\mathcal{G})$. We prove Proposition 5.2 by induction on the number of fixed marginals. Because the notation tends to be quite cumbersome, we will show that the Fréchet bounds for the cells in only the initial table $\mathbf{n}$ are attained. A similar argument can be made about every table in

$$\mathcal{RD}\setminus\bigcup_{r=1}^{p}\bigcup_{\{C:C\subseteq C_r\}}\mathcal{RD}(\mathbf{n}_C).$$

If $\mathcal{G}$ decomposes in $p=2$ cliques, we already proved that GSA converges to the Fréchet bounds in equation (12). We assume that Proposition 5.2 is true if $\mathbf{n}$ has at most $(p-1)$ fixed marginals that induce a decomposable independence graph. We want to prove Proposition 5.2 for an independence graph with p cliques. We take an arbitrary index $i^0\in\mathcal{I}$ that will remain fixed for the rest of this proof.

The cliques of $\mathcal{G}$ can be numbered so that they form a perfect sequence of vertex sets – see Lauritzen (1996). Let $H_{p-1}:=C_1\cup C_2\cup\ldots\cup C_{p-1}$. The subgraph $\mathcal{G}(H_{p-1})$ is decomposable and its cliques are $\{C_1,\ldots,C_{p-1}\}$, while its separators are $\{S_2,\ldots,S_{p-1}\}$. As before, $\mathbf{T}=\mathbf{T}^{(\mathbf{n})}$ is the set of cells associated with $\mathbf{n}$ defined in equation (4). In an analogous manner we define the set of cells $\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$ associated with the marginal table $\mathbf{n}_{H_{p-1}}$. The set of fixed cells $\mathbf{T}_0=\mathbf{T}_0^{(\mathbf{n})}\subset\mathbf{T}$ induced by fixing the cell counts in the marginals $\mathbf{n}_{C_1}$, $\mathbf{n}_{C_2}$, …, $\mathbf{n}_{C_p}$ of the table $\mathbf{n}$ includes the set of fixed cells $\mathbf{T}_0^{\left(\mathbf{n}_{H_{p-1}}\right)}\subset\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$ obtained by fixing the marginals $\mathbf{n}_{C_1},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_{p-1}}$ of the table $\mathbf{n}_{H_{p-1}}$.

We have $\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}\subset\mathbf{T}^{(\mathbf{n})}$ and $\mathcal{Q}\left(\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}\right)\subset\mathcal{Q}\left(\mathbf{T}^{(\mathbf{n})}\right)$. This implies that, when we run GSA for $\mathbf{T}^{(\mathbf{n})}$ and $\mathbf{T}_0^{(\mathbf{n})}$, it is as if we would run an instance ofGSA for $\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$ and $\mathbf{T}_0^{\left(\mathbf{n}_{H_{p-1}}\right)}$. Every vertex in $C_p\setminus S_p=C_p\setminus H_{p-1}$ is simplicial in the graph $\mathcal{G}$, hence Lemma 5.1 tells us that finding bounds for a cell in $t\in\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$ given $\mathbf{n}_{C_1},\mathbf{n}_{C_2},\ldots,\mathbf{n}_{C_{p-1}}$ is equivalent to finding

bounds for t given $\mathbf{n}_{C_1}, \mathbf{n}_{C_2}, \ldots, \mathbf{n}_{C_p}$. We do not lose any information by not considering the marginal $\mathbf{n}_{C_p}$ when computing bounds for $t \in \mathbf{T}\left(\mathbf{n}_{H_{p-1}}\right)$.

From the induction hypothesis we know that GSA employed for table $\mathbf{n}_{H_{p-1}}$ with the set of fixed cells $\mathbf{T}_0^{\left(\mathbf{n}_{H_{p-1}}\right)}$ converges to the Fréchet bounds for the cell $n_{H_{p-1}}\left(i^0_{H_{p-1}}\right)$:

$$\begin{aligned} n^U_{H_{p-1}}\left(i^0_{H_{p-1}}\right) &= \min\left\{n_{C_1}\left(i^0_{C_1}\right), \ldots, n_{C_{p-1}}\left(i^0_{C_{p-1}}\right)\right\}, \text{ and} \\ n^L_{H_{p-1}}\left(i^0_{H_{p-1}}\right) &= \max\left\{0, \sum_{r=1}^{p-1} n_{C_r}\left(i^0_{C_r}\right) - \sum_{r=2}^{p-1} n_{S_r}\left(i^0_{S_r}\right)\right\}. \end{aligned} \tag{35}$$

The shuttle procedure generates feasibility intervals $[L_s(t), U_s(t)]$ for every $t \in \mathbf{T}\left(\mathbf{n}_{H_{p-1}}\right)$. These are the tightest feasibility intervals GSA can find given the values of the cells in $\mathbf{T}_0^{\left(\mathbf{n}_{H_{p-1}}\right)}$. Because the information about the cells in the marginal $\mathbf{n}_{C_p}$ is not relevant for computing bounds for the cells in $\mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$, GSA employed for table $\mathbf{n}$ converges to the same feasibility intervals $[L_s(t), U_s(t)]$ for every $t \in \mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}$.

Since the sequence $C_1, C_2, \ldots, C_p$ is perfect in $\mathcal{G}$, $(H_{p-1} \setminus S_p, S_p, C_p \setminus S_p)$ is a proper decomposition of $\mathcal{G}$ – see Leimer (1993). Consider the graph $\mathcal{G}' = (K, E')$, where

$$E' := \{(u, v) : \{u, v\} \subset H_{p-1} \text{ or } \{u, v\} \subset C_p\}.$$

$\mathcal{G}'$ is a decomposable graph with two cliques H_{p-1}, C_p and one separator $H_{p-1} \cap C_p = S_p$. Running GSA for table $\mathbf{n}$ and the set of fixed cells $\mathbf{T}_0^{(\mathbf{n})}$ is equivalent to running GSA for $\mathbf{n}$ given the feasibility intervals $\left\{[L_s(t), U_s(t)] : t \in \mathbf{T}^{\left(\mathbf{n}_{H_{p-1}}\right)}\right\}$ and the set of fixed cells in $\mathbf{T}^{(\mathbf{n})}$ obtained by fixing the cells in the marginal $\mathbf{n}_{C_p}$.

As a consequence, by employing the shuttle procedure for table $\mathbf{n}$, we end up with the following Fréchet bounds for the count $n(i^0)$:

$$\begin{aligned} \min\left\{n^U_{H_{p-1}}\left(i^0_{H_{p-1}}\right), n_{C_p}\left(i^0_{C_p}\right)\right\} &\geq n(i^0), \text{ and} \\ n(i^0) \geq \max\left\{0, n^L_{H_{p-1}}\left(i^0_{H_{p-1}}\right) + n_{C_p}\left(i^0_{C_p}\right) - n_{S_p}\left(i^0_{S_p}\right)\right\}. \end{aligned} \tag{36}$$

It is straightforward to notice that equation (12) is obtained by combining equations (35) and (36). We can conclude that Proposition 5.2 is true when the set of fixed marginals are the minimal sufficient statistics of a decomposable log-linear model.

References

Bonferroni, C.E. (1936). "Teoria statistica delle classi e calcolo delle probabilitá." In *Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, Vol. 8, 1-62.

Buzzigoli, L. and Giusti, A. (1999). "An Algorithm to Calculate the Lower and Upper Bounds of the Elements of an Array Given its Marginals." In *Statistical Data Protection (SDP'98) Proceedings*, 131–147. Eurostat, Luxembourg.

Chen, Y., Dinwoodie, I. H., and Sullivant, S. (2006). "Sequential Importance Sampling for Multiway Tables." *The Annals of Statistics*, 34, 523-545.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. Springer Series in Statistics, Second ed. Springer-Verlag, New York.

Cox, L. H. (1999). "Some Remarks on Research Directions in Statistical Data Protection." In *Statistical Data Protection (SDP'98) Proceedings*, 163–176. Eurostat, Luxembourg.

Diaconis, P. and Sturmfels, B. (1998). "Algebraic algorithms for sampling from conditional distributions." *Annals of Statistics*, 26, 363–397.

Dobra, A. (2001). "Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables." Ph.D. thesis, Department of Statistics, Carnegie Mellon University.

— (2003). "Markov bases for decomposable graphical models." *Bernoulli*, 9, 6, 1–16.

Dobra, A. and Fienberg, S. E. (2000). "Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs." *Proceedings of the National Academy of Sciences*, 97, 11885–11892.

Dobra, A., Fienberg, S. E., Rinaldo, A., Slavkovic, A. B., and Zhou, Y. (2008). "Algebraic Statistics and Contingency Table Problems: Estimations and Disclosure Limitation." In *Emerging Applications of Algebraic Geometry*, eds. M. Putinar and S. Sullivant, IMA Series in Applied Mathematics. New York: Springer-Verlag.

Dobra, A., Fienberg, S. E., and Trottini, M. (2003a). "Assessing the Risk of Disclosure of Confidential Categorical Data." In *Bayesian Statistics 7*, eds. J. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 125–144. Oxford University Press.

Dobra, A., Karr, A., and Sanil, A. (2003b). "Preserving Confidentiality of High-dimensional Tabulated Data: Statistical and Computational Issues." *Statistics and Computing*, 13, 363–370.

Dobra, A. and Sullivant, S. (2004). "A Divide-and-conquer Algorithm for Generating Markov Bases of Multi-way Tables." *Computational Statistics*, 19, 347–366.

Dobra, A., Tebaldi, C., and West, M. (2006). "Data Augmentation in Multi-way Contingency Tables With Fixed Marginal Totals." *Journal of Statistical Planning and Inference*, 136, 355-372.

Edwards, D. E. (1992). "Linkage Analysis Using Log-linear Models." *Computational Statistics and Data Analysis*, 10, 281–290.

Edwards, D. E. and Havranek, T. (1985). "A fast procedure for model search in multidimensional contingency Tables." *Biometrika*, 72, 339–351.

Fienberg, S. E. (1999). "Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation." In *Statistical Data Protection (SDP'98) Proceedings*, 115–129. Eurostat, Luxembourg.

Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data, Second Edition*. MIT Press, Cambridge MA. Reprinted (2007) by Springer-Verlag, New York.

Fienberg, S. E. and Rinaldo, A. (2007). "Three Centuries of Categorical Data Analysis: Log-linear Models and Maximum Likelihood Estimation." *Journal of Statistical Planning and Inference*, 137, 3430–3445.

Fienberg, S. E. and Slavkovic, A. B. (2004). "Making the Release of Confidential Data from Multi-way Tables Count." *Chance*, 17, 5–10.

Fienberg, S. E. and Slavkovic, A. B. (2005). "Preserving the Confidentiality of Categorical Statistical Data Bases When Releasing Information for Association Rules." *Datamining and Knowledge Discovery*, 11, 155–180.

Fréchet, M. (1940). "Les Probabilitiés, Associées a un Système d'Événments Compatibles et Dépendants." Hermann & Cie, Paris, Vol. Premiere Partie.

Geiger, D., Meek, C., and Sturmfels, B. (2006). "On the Toric Algebra of Graphical Models." *The Annals of Statistics*, 34, 1463–1492.

Guo, S. W. and Thompson, E. A. (1992). "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles." *Biometrics*, 48, 361-372.

Haberman, S. J. (1978). *Analysis of Qualitative Data*. Academic Press, New York.

Hoeffding, W. (1940). "Scale-invariant Correlation Theory." In *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, Vol. 5(3), 181-233.

Hosten, S. and Sturmfels, B. (2007). "Computing the Integer Programming Gap." *Combinatorica*, 367–382.

Knuth, D. (1973). *The Art of Computer Programming*, vol. 3. Addison–Wesley.

Koch, G., Amara, J., Atkinson, S., and Stanish, W. (1983). "Overview of Categorical Analysis Methods." *SAS-SUGI*, 8, 785–795.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Leimer, H. G. (1993). "Optimal Decomposition by Clique Separators." *Discrete Mathematics*, 113, 99–123.

Manton, K. G., Corder, L., and Stallard, E. (1993). "Estimates of change in chronic disability and institutional incidence and prevalence rate in the U.S. elderly populations from 1982 to 1989." *Journal of Gerontology: Social Sciences*, 48, 153–166.

Sullivant, S. (2005). "Small Contingency Tables With Large Gaps." *SIAM Journal of Discrete Mathematics*, 18, 787–793.

Sundberg, R. (1975). "Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests." *Scandinavian Journal of Statistics*, 2, 71–79.

Vlach, M. (1986). "Conditions for the Existence of Solutions of the Three-dimensional Planar Transportation Problem." *Discrete Applied Mathematics*, 13, 61–78.

Whittaker, J.. (1990). *Graphical Models*. John Wiley & Sons.