Algebraic statistics and contingency tables

Adrian Dobra
University of Washington

AML08: Algebraic Methods in Machine Learning
Symposium and Workshop at NIPS'08

December 12, 2008

Some Revelant Publications

1. Dobra, A. and Fienberg, S.E. (2000). *Bounds for cell entries in contingency tables given marginal totals and decomposable graphs.* **PNAS**, 97(22), 1185–11892.
2. Dobra, A., Karr, A.F. and Sanil, A.P. (2003). *Preserving confidentiality of high-dimensional tabulated data: statistical and computational issues.* **Statistics and Computing**, 13, 363–370.
3. Dobra, A. (2003). *Markov bases for decomposable graphical models.* **Bernoulli**, 9(6), 1093–1108.
4. Dobra, A. and Sullivant, S. (2004). *A divide-and-conquer algorithm for generating Markov bases for multi-way tables.* **Computational Statistics**, 19, 347–366.
5. Dobra, A., Tebaldi, C. and West, M. (2006). *Data augmentation in multi-way contingency tables with fixed marginal totals.* **JSPI**, 136, 355–372.

Only 810 tables consistent with marginals $\mathcal{R}_1$!!.

$$\mathcal{R}_1 = \{[ACDEF], [ABDEF], [ABCDE], [BCDF], [ABCF], [BCEF]\}.$$

				B	no		yes		B	no		yes	
F	E	D	C	A	no	yes	no	yes	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67		[35, 45]	[35, 44]	[111, 121]	[63, 72]
			yes		129	145	12	23		[128, 138]	[141, 150]	[3, 13]	[18, 27]
		≥ 140	no		35	12	80	33		[29, 39]	[5, 14]	[76, 86]	[31, 40]
			yes		109	67	7	9		[105, 115]	[65, 74]	[1, 11]	[2, 11]
	≥ 3	< 140	no		23	32	70	66		[16, 25]	[26, 35]	[68, 77]	[63, 72]
			yes		50	80	7	13		[48, 57]	[77, 86]	[0, 9]	[7, 16]
		≥ 140	no		24	25	73	57		[19, 28]	[16, 25]	[69, 78]	[57, 66]
			yes		51	63	7	16		[47, 56]	[63, 72]	[2, 11]	[7, 16]
pos	< 3	< 140	no		5	7	21	9		[4, 14]	[3, 12]	[12, 22]	[4, 13]
			yes		9	17	$\boxed{1}$	4		[0, 10]	[12, 21]	$\boxed{[0, 10]}$	[0, 9]
		≥ 140	no		4	3	11	8		[0, 10]	[1, 10]	[5, 15]	[1, 10]
			yes		14	17	5	2		[8, 18]	[10, 19]	[1, 11]	[0, 9]
	≥ 3	< 140	no		7	3	14	14		[5, 14]	[0, 9]	[7, 16]	[8, 17]
			yes		9	16	2	3		[2, 11]	[10, 19]	[0, 9]	[0, 9]
		≥ 140	no		4	0	13	11		[0, 9]	[0, 9]	[8, 17]	[2, 11]
			yes		5	14	4	4		[0, 9]	[5, 14]	[0, 9]	[4, 13]

Table: Czeck Autoworkers data from Edwards & Havranek (1985) (left panel) and bounds given marginals $\mathcal{R}_1$ (right panel).

How to do inference under log-linear models $\mathcal{A}_1$–$\mathcal{A}_8$?

Log-linear Model	Minimal Sufficient Statistics
$\mathcal{A}_1$	$\mathcal{R}_1 \cup \{[BCDEF]\}$
$\mathcal{A}_2$	$\mathcal{R}_1 \cup \{[ABCEF]\}$
$\mathcal{A}_3$	$\mathcal{R}_1 \cup \{[ABCDF]\}$
$\mathcal{A}_4$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCEF]\}$
$\mathcal{A}_5$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCDF]\}$
$\mathcal{A}_6$	$\mathcal{R}_1 \cup \{[ABCEF], [ABCDF]\}$
$\mathcal{A}_7$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCEF], [ABCDF]\}$
$\mathcal{A}_8$	Saturated

$K = \{1, 2, \ldots, k\}$, $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ cross-classified in $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$.
$\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2 \times \ldots \times \mathcal{I}_k$, $\mathcal{I}_j = \{1, 2, \ldots, I_j\}$, $I_j \in \{1, 2, \ldots\}$.

Tables consistent with fixed marginals:

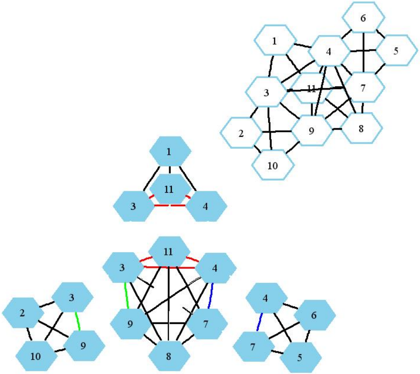
$$T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r}) = \{\mathbf{x} = \{x(i)\}_{i \in \mathcal{I}} : \mathbf{x}_{D_1} = \mathbf{n}_{D_1}, \ldots, \mathbf{x}_{D_r} = \mathbf{n}_{D_r}\}.$$

Questions of interest

1. Compute upper and lower bounds for cell entries:
$$\min\{\pm x(i) : i \in \mathcal{I}, \mathbf{x} \in T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})\}.$$
2. Enumerate tables in $T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$.
3. Estimate size of $T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$.
4. Sample from $T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$.
5. Probability distributions on $T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$.

- $G = (K, E)$ associated with $\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r}$ has edges:
$$E = \{(u, v) : \{u, v\} \subset D_j \text{ for some } j\}.$$
- Interpretation: if $(u, v) \notin E$, then
$$X_u \perp X_v | X_{K \setminus \{u,v\}} \Leftrightarrow u \perp v | K \setminus \{u, v\}.$$
- Special types of graphs:
  1. decomposable,
  2. reducible.

- $D_1 = \{1, 3, 4, 11\}$, $D_2 = \{3, 4, 7, 8, 9, 11\}$, $D_3 = \{2, 3, 9, 10\}$, $D_4 = \{4, 5, 6, 7\}$.
- $S_1 = \{3, 4, 11\}$, $S_2 = \{3, 9\}$, $S_3 = \{4, 7\}$.
- Fixed marginals: $\mathbf{n}_{D_1}$, $\mathbf{n}_{D_2}$, $\mathbf{n}_{D_3}$, $\mathbf{n}_{D_4}$.
- Cliques: $C(G) = \{D_1, D_2, D_3, D_4\}$; Separators: $S(G) = \{S_1, S_2, S_3\}$.



Theorem

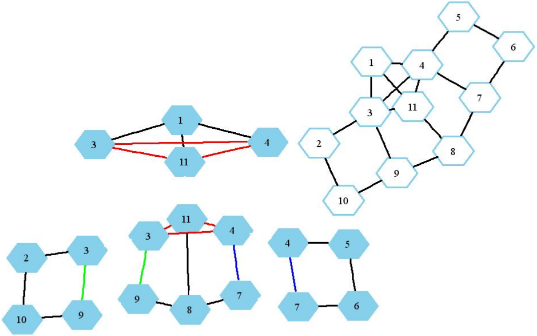
[Dobra & Fienberg, 2000] Let $G = (K, E)$ decomposable. Let $C(G)$ be the cliques of G and $S(G)$ the separators of G. Then:

$$\min\{n_C(i_C) | C \in C(G)\} \geq n(i) \geq$$
$$\max\left\{\sum_{C \in C(G)} n_C(i_C) - \sum_{S \in S(G)} n_S(i_S), 0\right\}.$$

Example:

$$\min\{n_{D_1}, n_{D_2}, n_{D_3}, n_{D_4}\} \geq n(i) \geq$$
$$\max\{n_{D_1} + n_{D_2} + n_{D_3} + n_{D_4} - n_{S_1} - n_{S_2} - n_{S_3}, 0\}.$$

- $D_1 = \{1,3,4,11\}$, $D_2 = \{3,4,7,8,9,11\}$, $D_3 = \{2,3,9,10\}$, $D_4 = \{4,5,6,7\}$.
- $S_1 = \{3,4,11\}$, $S_2 = \{3,9\}$, $S_3 = \{4,7\}$.
- Fixed marginals: $\mathbf{n}_{S_1}$ and all two-way marginals given by edges!!
- Prime components: $C(G) = \{D_1, D_2, D_3, D_4\}$; Separators: $S(G) = \{S_1, S_2, S_3\}$.



Theorem

[Dobra & Fienberg, 2000] Let $G = (K, E)$ reducible. Let $C(G)$ be the prime components of G and $S(G)$ the separators of G. Then:

$$\min\left\{n_C^U(i_C) | C \in C(G)\right\} \geq n(i) \geq$$
$$\max\left\{\sum_{C \in C(G)} n_C^L(i_C) - \sum_{S \in S(G)} n_S(i_S), 0\right\}.$$

Example:

$$\min\left\{n_{D_1}^U, n_{D_2}^U, n_{D_3}^U, n_{D_4}^U\right\} \geq n(i) \geq$$
$$\max\left\{n_{D_1}^L + n_{D_2}^L + n_{D_3}^L + n_{D_4}^L - n_{S_1} - n_{S_2} - n_{S_3}, 0\right\}.$$

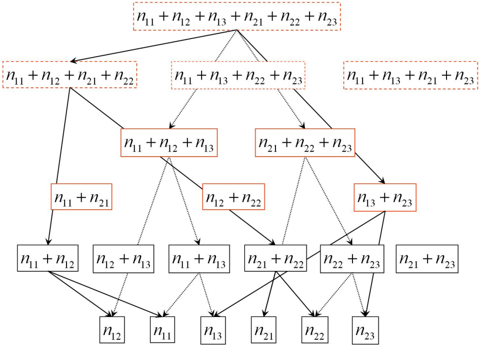
- Generalized version of the Shuttle Algorithm (Buzzigoli & Giusti).
- Exploit the tree-like structure of the problem.
- $\mathcal{C}$ cells obtained by collapsing across categories.
- New formulation of the bounds problem:

  Find the bounds $\mathcal{C}^U$ and $\mathcal{C}^L$ for the cells $\mathcal{C}$ given information about some cells $\mathcal{C}_0 \subset \mathcal{C}$.
- Let $c_1, c_2 \in \mathcal{C}$ such that their join c_{12} is still in $\mathcal{C}$. Then:

$$c_1^L + c_2^L \leq c_{12} \leq c_1^U + c_2^U,$$
$$c_{12}^L - c_2^U \leq c_1 \leq c_{12}^U - c_2^L.$$

Example: 2×3 table with fixed row and column totals.

$$
\begin{array}{ccc|c}
n_{11} & n_{12} & n_{13} & n_{1+} \\
n_{21} & n_{22} & n_{23} & n_{2+} \\
\hline
n_{+1} & n_{+2} & n_{+3} & n_{++}
\end{array}
$$



Calculating Cell Bounds

The generalized shuttle algorithm: Empty polytope!!

Example: There is no $6 \times 4 \times 3$ integer table having the two-way margins (Vlach, 1986):

$$n_{12} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \; n_{13} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \; n_{23} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Example: There are only two $3 \times 4 \times 6$ tables with marginals:

$$n_{12} = \begin{pmatrix} 2 & 2 & 2 & 2 \\ 3 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \end{pmatrix}, \; n_{13} = \begin{pmatrix} 2 & 1 & 2 & 3 & 0 & 0 \\ 2 & 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 2 & 1 & 2 & 3 \end{pmatrix},$$

$$n_{23} = \begin{pmatrix} 2 & 1 & 2 & 0 & 2 & 0 \\ 1 & 0 & 2 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 & 2 & 0 \\ 0 & 1 & 0 & 2 & 0 & 2 \end{pmatrix}.$$

Possible values for cell $(1, 1, 1)$ are 0 and 2!!

Example: 2^{16} table with three fixed 15-way marginals:

- $62,384$ zero entries out of $2^{16} = 65,536$ cells.
- Only 128 cells have upper bounds strictly bigger than lower bounds.
- $1,729$ (499) cells have counts of 1 (2).
- $1,698$ (485) of these cells have upper bounds equal lower bounds.

How to produce a full table consistent with a set of fixed marginals?

1. Global moves
   - Generated from the Generalized Shuttle Algorithm.
   - Could take a long time to compute.
   - Can balance between "long" and "short" jumps.
   - Can be used to estimate # of tables consistent with fixed marginals.
2. Local moves (Markov bases)
   - Formulas for decomposable case (Dobra, 2003).
   - Otherwise, need algebraic methods (Groebner bases).
   - Very fast once available!

- Order the cells in table: $\mathcal{I} = \{i^1, i^2, \ldots, i^m\}$.
- Possible current values for cell i^a:

$$\mathcal{H}_a := \{L(i^a), L(i^a) + 1, \ldots, U(i^a) - 1, U(i^a)\}.$$

- Choose scaling factors $v_a \in (0, 1)$, $a = 1, \ldots, m$.
- Generate a candidate table $\mathbf{n}^*$ as follows:
  - *for* $a = 1, \ldots, m$ *do*
    1. *Calculate current bounds* $L(i^a)$ *and* $U(i^a)$.
    2. *Draw a value* $n^*(i_a)$ *from* $\mathcal{H}_a$ *from proposal:*

$$q_a(n(i^a), n^*(i^a)) \propto v_a^{|n(i^a) - n^*(i^a)|}.$$

  end for

$\mathcal{M}(T)$ is number of tables in $T = T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$. Assume uniform distribution on T:

$$p(\mathbf{n}) = \frac{1}{\mathcal{M}(T)}.$$

Set scaling factors v_a equal to one:

$$q(n) \propto \prod_{a=1}^{m} \frac{1}{U(i^a) - L(i^a) + 1}.$$

Write:

$$1 = \sum_{\mathbf{n} \in T} \frac{p(\mathbf{n})}{q(\mathbf{n})} q(\mathbf{n}) \Rightarrow \mathcal{M}(T) = \sum_{\mathbf{n} \in T} \frac{1}{q(\mathbf{n})} q(\mathbf{n}).$$

Estimate $\hat{\mathcal{M}}(T) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{q(\mathbf{n}^{(S)})}$ where $\mathbf{n}^{(1)}, \ldots, \mathbf{n}^{(S)}$ independently sampled from $q(\cdot)$.

$\mathcal{R}_2$ are the 15 four-way marginals.

$\mathcal{R}_3 = \{[BF], [ABCE], [ADE]\}$

TABLE: Bounds given $\mathcal{R}_2$ (left-hand panel) and $\mathcal{R}_3$ (right-hand panel).

			B	no		yes		no		yes	
F	E	D	C \ A	no	yes	no	yes	no	yes	no	yes
neg	< 3	< 140	no	[27, 58]	[25, 56]	[96, 134]	[44, 82]	[0, 88]	[0, 62]	[0, 224]	[0, 117]
			yes	[108, 149]	[123, 168]	[0, 22]	[9, 37]	[0, 261]	[0, 246]	[0, 24]	[0, 38]
		≥ 140	no	[22, 49]	[0, 24]	[60, 96]	[16, 52]	[0, 88]	[0, 62]	[0, 224]	[0, 117]
			yes	[91, 127]	[45, 85]	[0, 18]	[0, 20]	[0, 261]	[0, 151]	[0, 24]	[0, 38]
	≥ 3	< 140	no	[10, 37]	[17, 44]	[48, 86]	[49, 89]	[0, 58]	[0, 60]	[0, 170]	[0, 148]
			yes	[30, 68]	[58, 102]	[0, 19]	[0, 25]	[0, 115]	[0, 173]	[0, 20]	[0, 36]
		≥ 140	no	[13, 37]	[8, 36]	[55, 90]	[38, 76]	[0, 58]	[0, 60]	[0, 170]	[0, 148]
			yes	[30, 67]	[45, 86]	[0, 19]	[0, 27]	[0, 115]	[0, 173]	[0, 20]	[0, 36]
pos	< 3	< 140	no	[0, 15]	[0, 13]	[4, 31]	[0, 23]	[0, 88]	[0, 62]	[0, 125]	[0, 117]
			yes	[0, 21]	[3, 30]	$\boxed{[0, 10]}$	[0, 9]	[0, 134]	[0, 134]	$\boxed{[0, 10]}$	[0, 38]
		≥ 140	no	[0, 11]	[0, 10]	[0, 24]	[0, 18]	[0, 88]	[0, 62]	[0, 125]	[0, 117]
			yes	[0, 26]	[2, 30]	[0, 11]	[0, 9]	[0, 134]	[0, 134]	[0, 24]	[0, 38]
	≥ 3	< 140	no	[1, 14]	[0, 9]	[0, 26]	[0, 26]	[0, 58]	[0, 60]	[0, 125]	[0, 125]
			yes	[0, 19]	[4, 29]	[0, 9]	[0, 9]	[0, 115]	[0, 134]	[0, 20]	[0, 36]
		≥ 140	no	[0, 9]	[0, 9]	[0, 26]	[0, 22]	[0, 58]	[0, 60]	[0, 125]	[0, 125]
			yes	[0, 19]	[0, 23]	[0, 9]	[0, 13]	[0, 115]	[0, 134]	[0, 20]	[0, 36]

Example: Czech Autoworkers

Estimating number of feasible table using global moves (II)

$\mathcal{R}_2$ are the 15 four-way marginals.
$705,884$ tables consistent with $\mathcal{R}_2$.
Estimated number of tables: $703,126$.
95% CI is $650,000$–$750,000$.

$\mathcal{R}_3 = \{[BF], [ABCE], [ADE]\}$.
Estimated number of tables consistent with $\mathcal{R}_3$: 10^{58}.
95% CI is 10^{57}–10^{59}.

Center for Statistics and the Social Sciences

DEFINITION

A local move $\mathbf{g} = \{g(i)\}_{i \in \mathcal{I}}$ is a multi-way array with integer entries $g(i) \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.

DEFINITION

A Markov basis for $T = T(\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r})$ allows any two tables $\mathbf{n}_1$, $\mathbf{n}_2$ in T to be connected by a series of local moves:

$$\mathbf{n}_1 - \mathbf{n}_2 = \sum_{j=1}^{r} \mathbf{g}^j.$$

Primitive moves for 2x2 tables:

$$
\begin{array}{cc|c}
n_{11}+1 & n_{12}-1 & n_{1+} \\
n_{21}-1 & n_{22}+1 & n_{2+} \\
\hline
n_{+1} & n_{+2} & n_{++}
\end{array}
$$

Primitive moves for two-way tables:

$$
g^{i_1 i_2; j_1 j_2}(i,j) = \left\{ \begin{array}{cc} 1, & \text{if } (i,j) \in \{(i_1,j_1),(i_2,j_2)\}. \\ -1, & \text{if } (i,j) \in \{(i_1,j_2),(i_1,j_2)\}, \\ 0, & \text{otherwise.} \end{array} \right. .
$$

Extension to decomposable models with two cliques!!

Center for Statistics and the Social Sciences

THEOREM

[Dobra, 2001] A well defined set of primitive moves connects all tables having a set of fixed marginals $\mathbf{n}_{D_1}, \ldots, \mathbf{n}_{D_r}$*, when this set of marginals are the cliques* $\{D_1, \ldots, D_r\}$ *of a decomposable graph* $G = (K, E)$*.*

PROOF.

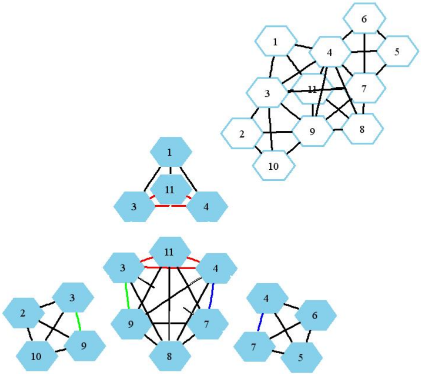
For every separator S_j of G there exists a proper decomposition of G: $(V_j^1 \setminus S_j, S_j, V_j^2 \setminus S_j)$. A Markov basis for $D_1, \ldots, D_r$ is:

$$\mathrm{MB}(D_1, \ldots, D_r) = \bigcup_{j=2}^{r} \mathrm{F}(V_j^1, V_2^j).$$

$\square$

Divide-and-conquer technique to generate Markov bases for reducible graphs!!

- $D_1 = \{1,3,4,11\}$, $D_2 = \{3,4,7,8,9,11\}$, $D_3 = \{2,3,9,10\}$, $D_4 = \{4,5,6,7\}$.
- $S_1 = \{3,4,11\}$, $S_2 = \{3,9\}$, $S_3 = \{4,7\}$.
- $\mathrm{MB}(D_1, D_2, D_3, D_4) = \mathrm{F}(D_1, \{2,\ldots,11\}) \cup \mathrm{F}(D_2, \{1,3,\ldots,9,11\}) \cup \mathrm{F}(D_4, \{1,\ldots,4,7,\ldots,11\})$.



- $\mathbf{n} = \{n(i)\}_{i \in \mathcal{I}}$ multi-way table.
- $\mathcal{D}$ is available data (e.g., marginals, bounds, structural zeros).
- $\mathcal{T}$ tables consistent with $\mathcal{D}$.
- Cell counts $n(i) \sim Poisson(\lambda(i))$, $\lambda(i) > 0$.

$$
\begin{aligned}
p(\mathcal{D}|\lambda) &= \sum_{\mathbf{n}' \in \mathcal{T}} p(\mathbf{n}'|\lambda). \\
p(\mathbf{n}, \mathcal{D}|\lambda) &= p(\mathbf{n}|\lambda) \cdot I_{\{\mathbf{n} \in \mathcal{T}\}}. \\
p(\mathbf{n}|\mathcal{D}, \lambda) &= \frac{p(\mathbf{n}|\lambda)}{p(\mathcal{D}|\lambda)} \cdot I_{\{\mathbf{n} \in \mathcal{T}\}}.
\end{aligned}
$$

Let $\mathcal{A}$ log-linear model for $\lambda = \{\lambda(i)\}_{i \in \mathcal{I}}$:

$$\lambda(i) = \mu \prod_{C} \psi_C(i_C).$$

Theorem

If marginal $\mathbf{n}_C$ determined from $\mathcal{D}$, then, under $\mathcal{A}$, $p(\mathbf{n}|\mathcal{D}, \lambda)$ does not depend on $\psi_C(i_C)$.

Theorem

The hypergeometric distribution is obtained by conditioning on a log-linear model whose parameters are determined from $\mathcal{D}$.

Example: Czech Autoworkers
Cell Bounds and Table Counting

Only 810 tables consistent with marginals $\mathcal{R}_1$!!.

$$\mathcal{R}_1 = \{[ACDEF], [ABDEF], [ABCDE], [BCDF], [ABCF], [BCEF]\}.$$

				B	no		yes		B	no		yes	
F	E	D	C	A	no	yes	no	yes	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67		[35, 45]	[35, 44]	[111, 121]	[63, 72]
			yes		129	145	12	23		[128, 138]	[141, 150]	[3, 13]	[18, 27]
		≥ 140	no		35	12	80	33		[29, 39]	[5, 14]	[76, 86]	[31, 40]
			yes		109	67	7	9		[105, 115]	[65, 74]	[1, 11]	[2, 11]
	≥ 3	< 140	no		23	32	70	66		[16, 25]	[26, 35]	[68, 77]	[63, 72]
			yes		50	80	7	13		[48, 57]	[77, 86]	[0, 9]	[7, 16]
		≥ 140	no		24	25	73	57		[19, 28]	[16, 25]	[69, 78]	[57, 66]
			yes		51	63	7	16		[47, 56]	[63, 72]	[2, 11]	[7, 16]
pos	< 3	< 140	no		5	7	21	9		[4, 14]	[3, 12]	[12, 22]	[4, 13]
			yes		9	17	$\boxed{1}$	4		[0, 10]	[12, 21]	$\boxed{[0, 10]}$	[0, 9]
		≥ 140	no		4	3	11	8		[0, 10]	[1, 10]	[5, 15]	[1, 10]
			yes		14	17	5	2		[8, 18]	[10, 19]	[1, 11]	[0, 9]
	≥ 3	< 140	no		7	3	14	14		[5, 14]	[0, 9]	[7, 16]	[8, 17]
			yes		9	16	2	3		[2, 11]	[10, 19]	[0, 9]	[0, 9]
		≥ 140	no		4	0	13	11		[0, 9]	[0, 9]	[8, 17]	[2, 11]
			yes		5	14	4	4		[0, 9]	[5, 14]	[0, 9]	[4, 13]

Table: Czeck Autoworkers data from Edwards & Havranek (1985) (left panel) and bounds given marginals $\mathcal{R}_1$ (right panel).

How to do inference under log-linear models $\mathcal{A}_1$–$\mathcal{A}_8$?

Log-linear Model	Minimal Sufficient Statistics
$\mathcal{A}_1$	$\mathcal{R}_1 \cup \{[BCDEF]\}$
$\mathcal{A}_2$	$\mathcal{R}_1 \cup \{[ABCEF]\}$
$\mathcal{A}_3$	$\mathcal{R}_1 \cup \{[ABCDF]\}$
$\mathcal{A}_4$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCEF]\}$
$\mathcal{A}_5$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCDF]\}$
$\mathcal{A}_6$	$\mathcal{R}_1 \cup \{[ABCEF], [ABCDF]\}$
$\mathcal{A}_7$	$\mathcal{R}_1 \cup \{[BCDEF], [ABCEF], [ABCDF]\}$
$\mathcal{A}_8$	Saturated

Let $\mathcal{A}$ log-linear model with parameters θ for Poisson means λ.

At the s-th step of algorithm do:

1. Simulate $\theta^{(s+1)}$ from $p(\theta|\mathcal{A}, \mathbf{n}^{(s)}) \propto p(\theta|\mathcal{A})p(\mathbf{n}^{(s)}|\lambda(\theta))$. Compute $\lambda^{(s+1)} = \lambda(\theta^{(s+1)})$.
2. Simulate $\mathbf{n}^{(s+1)}$ from $p(\mathbf{n}|\mathcal{D}, \lambda^{(s+1)})$.

Independent Gamma (Uniform) priors for λ imply conjugate (truncated) Gamma posteriors (West, 1997; Tebaldi & West (1998)).

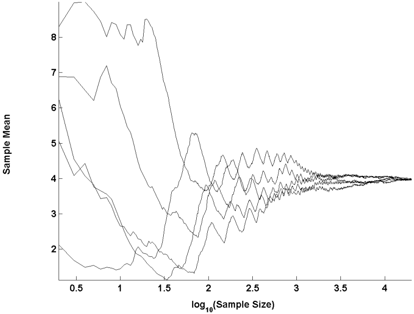


Figure: Convergence of the data augmentation method for the Czech autoworkers data. The x-axis represents the iteration number on a $\log_{10}$ scale, while the *y*-axis gives the sample mean of λ_0 from five starting points under model $\mathcal{A}_8$.

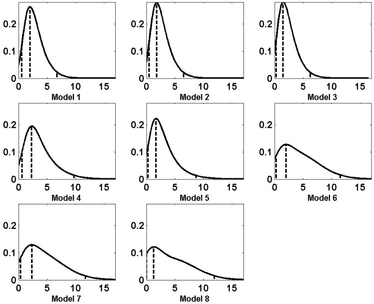


Figure: Approximate posterior distributions for λ_0 under the log-linear models $\mathcal{A}_1, \ldots, \mathcal{A}_8$. The dotted lines represent estimates of the posterior mode and the corresponding 95% confidence intervals.

Next Steps...

(I) Computing exact p-values.

(II) Methods for sparse tables.