Conditional Inference given Partial Information for Contingency Tables

Aleksandra Slavković

Department of Statistics
Pennsylvania State University

AMS Alg. Stats. Session
March 27, 2010

Acknowledgments

- Juyoun Lee, Vishesh Karwa, Sonja Petrović, Xiaotian Zhu
- NSF grant SES-0532407 to the Department of Statistics, Penn State University

- References for papers and *R* code:
  - http://www.stat.psu.edu/~sesa/
  - http://www.stat.psu.edu/~sesa/cctable

What can we learn from fragmentary (but compatible) data?

- Given some (partial) information ($\mathbf{T}$)that is related to unobserved contingency table ($\mathbf{n}$), what can we learn about that table and its joint distribution ($\mathbf{p}$)?
- What reliable statistical analysis is possible?, e.g., $f(\mathbf{n}, \mathbf{p}|\mathbf{T}) \approx f(\mathbf{n}, \mathbf{p}|\mathit{full})$
- Conditional inference given partial information: optimization, enumeration, sampling.

Relevant for data privacy and confidentiality, ecological inference, missing data problems, causal inference with observational data.

Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) & contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal totals, conditional rates

Example

Table: A 2x2x2 Table on illegal MP3 downloading

Building	Gender	Download Yes	Download No	Total
A	Male	8	4	12
A	Female	2	9	11
B	Male	7	6	13
B	Female	3	11	14
	Total	20	30	50

What can we learn about the table **n** and **p**?

Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) & contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal totals, conditional rates

Example

Table: [Gender, Download] Marginal table of illegal MP3 downloading

Gender	Download Yes	No	Total
Male	15	10	25
Female	5	20	25
Total	20	30	50

What can we learn about the table **n** and **p**?

Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) & contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal totals, conditional rates

Example

Table: [Download|Gender] Table of conditional probabilities with [rounded probability]

Gender	Download Yes	Download No	Total
Male	$\frac{15}{25} = \frac{3}{5}$ [0.6]	$\frac{10}{25} = \frac{2}{5}$ [0.4]	25
Female	$\frac{5}{25} = \frac{1}{5}$ [0.2]	$\frac{20}{25} = \frac{4}{5}$ [0.8]	25
Total	20	30	50

What can we learn about the table **n** and **p**?

Conditional Inference: sampling with Markov bases

$\mathbf{n} \sim \mathit{Mutlinomial}(n, \mathbf{p})$
$\mathbf{p} \in \triangle = \{\mathbf{p} : \mathbf{p}(i) \geq 0 \text{ and } \sum_{i \in \mathcal{I}} \mathbf{p}(i) = 1 \ \forall i, j\}$
$\mathbf{T}$ is the given partial information, i.e., linear constraints
$\mathcal{F}_{\mathbf{T}}$ the set of all possible tables that preserve $\mathbf{T}$

$$\mathcal{F}_{\mathcal{T}} = M^{-1}[\mathbf{t}] := \{\mathbf{n} \in \mathbb{Z}_+^d : M\mathbf{n} = \mathbf{t}\}$$

MCMC methods to compute expectation of $f(\mathbf{n}, \mathbf{p})$ given $\mathbf{T}$

$$P(\mathbf{N} = \mathbf{n} | \mathbf{N} \in \mathcal{F}_{\mathbf{T}}) = \int_{\triangle} P(\mathbf{n}, \mathbf{p} | \mathcal{F}_{\mathbf{T}}) d\mathbf{p} = \int_{\triangle} P(\mathbf{n} | \mathcal{F}_{\mathbf{T}}, \mathbf{p}) \pi(\mathbf{p}) d\mathbf{p},$$

$$P(\mathbf{n} | \mathcal{F}_{\mathbf{T}}, \mathbf{p}) = \frac{P(\mathbf{n}|\mathbf{p}) I_{\mathbf{n} \in \mathcal{F}_{\mathbf{T}}}}{\sum_{\mathbf{n}' \in \mathcal{F}_{\mathbf{T}}} P(\mathbf{n}'|\mathbf{p})}, \text{ where } P(\mathbf{n}, \mathbf{T} | \mathbf{p}) = \begin{cases} P(\mathbf{n}|\mathbf{p}), & \mathbf{n} \in \mathcal{F}_{\mathbf{T}} \\ 0, & \mathit{otherwise}. \end{cases}$$

Algebraic Algorithms for Generating Tables

- When $\mathbf{T}$ is a set of marginal totals
  Diaconis and Sturmfels (1998), Dobra et al (2006), Chen et al (2006)
  - Hypergeometric distribution is a special well-known case of $P(\mathbf{n}|\mathbf{T}, \mathbf{p}) = P(\mathbf{n}|\mathbf{T})$.

- When $\mathbf{T}$ is a set of conditional rates & N:
  Slakovic (2004), Lee(2009), Slakovic & Lee (2009)
  - Prior and posterior specification of $\mathbf{p}$
  - Generate a *synthetic table*.
  - Understand the structure of $\mathcal{F}_{\mathbf{T}}$, i.e. support

- We explore the connections between $\mathcal{F}_{cond}$ & $\mathcal{F}_{marg}$.

  [Dobra et al. 2008] *"Problem 5.7. Characterize difference of two fibers, one for a conditional probability array, and the other for the corresponding margin, and thus simplify the calculation of Markov bases for the conditionals by using the knowledge of the moves of the corresponding margins."*

Problem Statement: Notation

Consider k categorical random variables, $X_1, \dots, X_k$, where each X_i takes value on the finite set of categories $[d_i] \equiv \{1, \dots, d_i\}$.

The cross-classification of N iid realizations of $(X_1, \dots, X_k)$ produces a random integer-valued array $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, a k-way *contingency table*.

Let A, B be nonempty and $A \cup B$ proper subset of $\{X_1, X_2, ..., X_k\}$.

Let $C = \{X_1, X_2, ..., X_k\} \setminus (A \cup B)$.

Summarize $\mathbf{n}$ as a 3-way table $\mathbf{n}^* = \{\mathbf{s_{ijk}}\}$ where s_{ijk} is the count in the cell: $A = i, B = j, C = k$.

Also let c_{ij} be the conditional frequency $P(A = i | B = j)$, and suppose it is equal $\frac{g_{ij}}{h_{ij}}$ where g_{ij} and h_{ij} are nonnegative integers and are relatively prime.

Problem statement: Space of tables

Investigate the space of all possible tables $\mathbf{n}$ consistent with:

(a) the grand total, $\sum_{i_1 \ldots i_k} n_{i_1 i_2 \ldots i_k}$, is N.

(b) a set of conditional frequencies, $P(A|B)$

- These can be either *full* or *small* conditionals.
- All of the given frequencies are exact.
- $\mathcal{F}_{A|B,N}$ can be expressed as the integer solutions of

$$\left\{ \begin{array}{l} \mathbf{Mn} = \mathbf{t} \\ \text{every B marginal} > 0 \end{array} \right\} \tag{1}$$

  where $\mathbf{n}$ and t are length d column vectors, $\mathbf{M}$ is a $J+1$ by d matrix
- Explore the links between $\mathcal{F}_{A|B}$ & $\mathcal{F}_{AB}$: implications for bounds and Markov bases.

Link between conditionals and marginals: Linear Diophantine equation

Theorem

Let m_j be the least common multiple of all h_{ij} for fixed j, and let $J = |B|$, the number of values that B takes. Then, each positive integer solution $\{x_j\}_{j=1}^{J}$ of

$$\sum_{j=1}^{k} m_j \cdot x_j = N \tag{2}$$

corresponds to a marginal s_{+j+}, up to a scalar multiple m_j. In particular, a table $\mathbf{n}$ consistent with the given information $\{c_{ij}, N\}$ exists if and only if Equation (2) has a positive integer solutions.

- Each solution of (2) corresponds to a marginal:, i.e., $s_{+j+} = m_j x_j$.
- The marginal determines the exact (integer) cell bounds of $\mathbf{n}$, i.e., . $[0, s_{+j+} \cdot d_{ij}]$.
- A different marginal $\{s_{+j+}\}$ certainly leads to different cell bounds.

Link between conditionals and marginals

Corollary

Let $\mathcal{F}_{A|B}$ be the space of tables given $\mathcal{T} = \{P(A|B), N\}$, where we allow for full conditionals. In addition, let $\mathcal{F}_{AB}$ and the space of tables given the corresponding $[AB]$ marginal counts s_{ij+}. Then, the following statements are equivalent:

(a) *$\mathcal{F}_{A|B}$ coincides with $\mathcal{F}_{AB}$.*

(b) *Equation (2) has only one positive integer solution.*

- The integer cell bounds are same: $0 \leq n_{ijk} \leq s_{ij+} = m_j x_j c_{ij}$
- *solvequick()* function in R to find the number of solutions to (2).

Link between conditionals and marginals: Table-space decomposition result

Corollary

Let $\mathcal{F}_{A|B}$ be the space of tables given $\mathcal{T} = \{P(A|B), N\}$, where we allow for full conditionals. Suppose that the Diophantine equation (2) has m solutions. Denote by $\mathfrak{p}_i$ the marginal corresponding to the i^{th} solution. Thus, we will denote the space of tables given that particular marginal table by $\mathcal{F}_{AB}(\mathfrak{p}_i)$. Then, we have the following decomposition of the table space taken as a disjoint union:

$$\mathcal{F}_{A|B} = \bigcup_{i=1}^{m} \mathcal{F}_{AB}(\mathfrak{p}_i).$$

Consequence for counting: exact and approximate

Lemma (Exact count of data tables given one marginal)

Corollary (Exact count of data tables given conditionals)

The number of possible k-way tables given observed conditionals $[A|B]$ *is*

$$|\mathcal{F}_{A|B}| = \sum_{i=1}^{m} |\mathcal{F}_{AB}(\mathfrak{p}_i)|,$$

where $\mathcal{F}_{AB}(\mathfrak{p}_i)$ *is as defined in Corollary 4, and m is the number of integer solutions to (2). Each* $|\mathcal{F}_{AB}(\mathfrak{p}_i)|$ *can be computed using Lemma* **??**.

Proposition (Approximate count of marginal tables given conditionals)

Corollary (Approximate count of data tables given conditionals)

Bounds on & counting of $[Building, Gender, Download]$

- Suppose we release $N = 50$ and a small conditional $P(Download|Gender)$:

$$\begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

Then equation (2) for this example is: $5x_1 + 5x_2 = 50$

Example

Table: A 2x2x2 Table on illegal MP3 downloading, & bounds

Building	Gender	Download Yes	No	Total
A	Male	8 [0,29.4] [0,27] [0,15]	4 [0,19.6] [0,18] [0.10]	12
A	Female	2 [0,9.8] [0,9] [0,5]	9 [0,39.2] [0,36] [0,20]	11
B	Male	7 [0,29.4] [0,27] [0,15]	6 [0,19.6] [0,18] [0.10]	13
B	Female	3 [0,9.8] [0,9] [0,5]	11 [0,39.2] [0,36] [0,20]	14
	Total	20	30	50

- 9 positive integer solutions: $\{(x_1 = i, x_2 = 10 - i) | 1 \leq i \leq 9\}$.
- 9 different [Download,Gender] marginals
- $|\mathcal{F}_{D|G}| = \sum_i |\mathcal{F}_{DG_i}| = 129778$
- $|\mathcal{F}_{D|G}| \geq |\mathcal{F}_{DG}|$ => release conditionals

Consequence for Markov bases and sampling

Corollary

The Markov basis for the space of tables given the conditional can be split into two sets of moves:

1) *the set of moves that fix the margin, and*
2) *the set of moves that change the margin.*

The Markov basis connecting all of $\mathcal{F}_{A|B}$ consists of the moves connecting each sub-fiber $\mathcal{F}_{AB}(\mathfrak{p}_i)$ (the first set of moves) and the moves connecting each sub-fiber to another (the second set of moves).

Computations suggest the following:

Corollary (Conjecture)

A minimal Markov basis of M in (1) contains $|B| - 1 + (|C| - 1) \times |B| \times |A|$ *elements.*

Reference Set for Marginal versus Conditional

Markov moves given the $[D|G] == [A|B]$ are

0	0	0	1	0	0	0	-1
0	0	1	0	0	0	-1	0
0	1	0	0	0	-1	0	0
1	0	0	0	-1	0	0	0
3	2	-1	-4	0	0	0	0

- Perturbation using first four moves are maintaining $[AB]$.
- Perturbation using the last moves are varying $[AB]$.
  - Last move based on rounded $[A|B]$ allows more than one possible values for $[AB]$.
  - $\mathcal{F}_{[A|B]} = \mathcal{F}_{AB_1} \biguplus \cdots \biguplus \mathcal{F}_{AB_l}$

Reference set given conditionals, for example $[A|B]$, is corresponding to either

- reference set given the corresponding marginal $[AB]$
- disjoint union of finite reference sets given $[AB]$

Sampling tables given $[A|B]$ = Combining samplings given $[AB]'s$

MCMC: Algorithm 1

1. Sample $\mathbf{p}^{(t+1)}$ from $P(\mathbf{p}|\mathbf{n}^{(t)}, \mathbf{T}) \propto P(\mathbf{n}^{(t)}|\mathbf{p})P(\mathbf{p}|\mathbf{T}) = P(\mathbf{n}^{(t)}|\mathbf{p})P(\mathbf{p})$. For example, the prior density for $\mathbf{p}$ is assumed to be Dirichlet distribution with hyper-parameters, $\eta = \{\eta(i)\}$ then $P(\mathbf{p}|\mathbf{n}^{(t)}, \mathbf{T})$ is proportional to Dirichlet with $\mathbf{n}^{(t)} + \eta$. $\mathbf{p}^{(t+1)}$ is drawn from Dirichlet distribution.
2. Generate tables from the conditional distribution, $P(\mathbf{n}|\mathbf{T}, \mathbf{p})$, is divided into two steps: completing a table consistent with the given information and deciding to accept or reject it.
    1. Generate the candidate table $\mathbf{n}^*$ from $q(\mathbf{n}^{(t)}, \mathbf{n}^*)$ induced by Markov moves. Uniformly choose one move $\mathbf{m} \in MB$ and $\epsilon = \pm 1$ with equal probability
    2. Add the selected move to the previous table, that is, $\mathbf{n}^* = \mathbf{n}^{(t)} + \epsilon \mathbf{m}$.
3. If $\mathbf{n}^* \geq 0$, accept the candidate table $\mathbf{n}^*$ with $\min\{1, \rho\}$, where

$$\rho = \frac{P(\mathbf{n}^*|\mathbf{p}^{(t)})}{P(\mathbf{n}^{(t)}|\mathbf{p}^{(t)})}. \tag{3}$$

Otherwise, stay at $\mathbf{n}^{(t)}$.

MCMC: Algorithm 2

1. For $l = 1, \ldots, L$, simulate contingency tables, $\mathbf{n}_{l,1}, \ldots, \mathbf{n}_{l,S_l}$ from the sub-reference set, $\mathcal{F}_{AB^l}$ or $\mathcal{F}_{AB^l,C}$ via a certain sampling scheme, for example, the MCMC with algebraic tools in [2], [3], and [?].
2. Average/Combine L sets of sampled tables.

$$P(\mathbf{N} = \mathbf{n} | \mathbf{n}_{A|B}, n) = \sum_{l=1}^{L} P(\mathbf{N} = \mathbf{n} | \mathbf{n}_{AB^l}, n) w_l, \tag{4}$$

where $w_l = P(\mathbf{n}_{AB^l} | \mathbf{n}_{A|B}, n)$, and $\mathbf{n}_{AB^l}$ is consistent with $\mathbf{n}_{A|B}$ for $l = 1, \ldots, L$

- Assigning Weights 1: Equal Weights
  $w = w_1 = \ldots = w_L$.

$$P(\mathbf{N} = \mathbf{n} | \mathbf{n}_{A|B}, n) = w \sum_{l=1}^{L} P(\mathbf{N} = \mathbf{n} | \mathbf{n}_{AB^l}, n). \tag{5}$$

- Assigning Weights 2: Markov Moves Assign more weight on the sub-reference set preserving the original values for the marginal $[AB]$. $w_1 = \frac{|MB_{AB}|}{|MB_{A|B}|}$ and $w_i = \frac{1}{L-1} \frac{|MB_{AB'}|}{|MB_{A|B}|}$ for $i = 2, \ldots, L$.

Current and Future Work

- Other sampling schemes [Lee (2009)]
  - Algorithm 3: Combination of multiple MCMC samplers
  - Algorithm 4: Importance sampling
- Rounding issues with conditional rates
- Prior and Posterior specification on λ
- Implications for Bounds
  - multiple conditionals
  - combination of marginals and conditionals, e.g. link to DAGs [Slavkovic, Zhu and Petrovic (under rev.)], & GSA
- Synthetic Tables [Slavkovic & Lee (2009)]
- Applications to ecological inference and causal inference with observational data [Karwa and Slavkovic (in prep)].
- Implementations in R, & interfacing with 4ti2.

References

Chen, Y. and Dinwoodie, I.H. and Sullivant, S. (2006)
Sequential importance sampling for multiway tables. *Ann. Statist*

Diaconis, P. and Sturmfels, B. (1998).
Algebraic algorithms for sampling from conditional distributions. *Ann. Statist*

Dobra, A. and Tebaldi, C. and West, M. (2006).
Data augmentation in multi-way contingency tables with fixed marginal totals. *J of Stat. Planning and Inference*

Dobra, A. and Fienberg, SE. and Rinaldo, A. and Slavkovic, A. and Zhou, Y.
Algebraic statistics and contingency table problems: Estimations and disclosure limitation. *Emerging Applications of Algebraic Geometry 2008*.

Slavković, A. & Lee (2009).
Synthetic tabular data preserving observed conditional probabilities. *Stat. Meth.*.

Lee, J. and Slavković, A. (2008).
Posterior distributions for the unobserved cell counts in contingency tables. *ISBA 2008*.

Lee, J. (2009).
Sampling contingency tables given marg./cond. *PSU thesis*.

Slavkovic, A. and Petrovic, S. and Zhu, X. (2009)
Mathematical Aspects of Space of Confidential Contingency Tables. *under rev.*.

Thank you.