Applied Algebraic Statistics Framework for Causal Inference

Vishesh Karwa, Aleksandra Slavković

Department of Statistics
Pennsylvania State University

SIAM Algebraic Statistics Session
July 13, 2010

Acknowledgments

- NSF grant SES-0532407 to the Department of Statistics, Penn State University

Outline

Motivation

Consider a study where the causal effect of T (e.g. smoking) on Y (e.g. lung cancer) is of interest.
Assume that the data is generated by the following (qualitative) model:

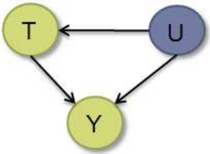

Figure: Causal Model

Algebraic statistics & Inference from partial data

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.

Conditional Inference: sampling with Markov bases

$X_1, \ldots, X_k$, where each $X_i \in [d_i] \equiv \{1, \ldots, d_i\}$
$\mathbf{n} \sim$ multinomial$(N, \mathbf{p})$
$\mathbf{p} \in \triangle = \{\mathbf{p} : \mathbf{p}(i) \geq 0 \ \forall i$ and $\sum_{i \in \mathcal{I}} \mathbf{p}(i) = 1 \}$

$\mathcal{M}$ is a statistical model specified by a set of (semi-algebraic) constraints on $\mathbf{p}$
$\mathbf{T}$ is the given partial information, i.e., linear constraints
$\mathcal{F}_{\mathbf{T}}$ the set of all possible tables that preserve $\mathbf{T}$

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

A is the constraint matrix:

Example

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ d_{12} & -d_{11} & 0 & 0 \\ 0 & 0 & d_{22} & -d_{21} \end{pmatrix}, \mathbf{t} = \begin{pmatrix} N \\ n_{1+} \\ n_{2+} \\ 0 \\ 0 \end{pmatrix}.$$

Algebraic Algorithms for Conditional Inference

- When **T** is a set of marginal totals
  - Diaconis and Sturmfels (1998) - Markov Basis and log-linear models
  - Dobra et al (2006), Chen et al (2006)

- When **T** is a set of conditional rates & N:
  - Slakovic (2004) - Generate a *synthetic table*.
  - Lee(2009), Slakovic & Lee (2009) - Prior and posterior specification of **p**

- When **T** is a set of arbitrary linear constraints:
  - Marginals, Conditional rates, population zeros etc.
  - Useful for several applied problems in statistical inference

Basic MCMC algorithm

- Straight forward extension of MCMC algorithm in Diaconis and Sturmfels (1998)
- When $\mathbf{T} = \{\text{margins}\}$ and $\mathcal{M} = \text{log-linear}$:
  - $\mathbf{T}$ is *MSS*
  - $P(\mathbf{n}|\mathbf{T}, \mathcal{M})$ does not depend on $\mathbf{p}$
- In general
  - $\mathbf{T}$ need not be *MSS* of $\mathcal{M}$
  - $P(\mathbf{n}|\mathbf{T}, \mathcal{M})$ depends on $\mathbf{p}$

Sample from $P(\mathbf{n}, \mathbf{p}| T, \mathcal{M})$, Use Variable at a time MCMC:

- Sample from $P(\mathbf{n}|\mathbf{p}, \mathbf{T}, \mathcal{M})$
- Sample from $P(\mathbf{p}|\mathbf{n}, \mathbf{T}, \mathcal{M})$

R4ti2 - interface with 4ti2

- R interface to 4ti2, R4ti2 [Karwa and Slavkovic (in prep.)]
  - constraint()
  - markovBasis()
  - groebnerBasis()
  - mcmc1()
  - mcmc2()
  - pvalue(), ecological() etc.

MCMC: Algorithm 1

1. Sample $\mathbf{p}^{(t+1)}$ from $P(\mathbf{p}|\mathbf{n}^{(t)}, \mathbf{T}, \mathcal{M}) \propto P(\mathbf{n}^{(t)}|\mathbf{p})P(\mathbf{p}|\mathbf{T}, \mathcal{M}) = P(\mathbf{n}^{(t)}|\mathbf{p})P(\mathbf{p}|\mathcal{M})$. (Could be a Gibbs update, e.g for multinomial with Dirichlet distribution or may require M-H sampling for non-standard distributions)
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

$$\rho = \frac{P(\mathbf{n}^*|\mathbf{p}^{(t)})}{P(\mathbf{n}^{(t)}|\mathbf{p}^{(t)})}. \tag{1}$$

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

MCMC: Algorithm 2

Algorithm 2 based on the following corollary due to Slavkovic, Zhu, and Petrovic (2009)

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).

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
2. Average/Combine L sets of sampled tables.

$$P(\mathbf{N} = \mathbf{n}|\mathbf{n}_{A|B}, n) = \sum_{l=1}^{L} P(\mathbf{N} = \mathbf{n}|\mathbf{n}_{AB^l}, n) w_l, \quad (2)$$

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
  $w = w_1 = \ldots = w_L$.

$$P(\mathbf{N} = \mathbf{n}|\mathbf{n}_{A|B}, n) = w \sum_{l=1}^{L} P(\mathbf{N} = \mathbf{n}|\mathbf{n}_{AB^l}, n). \quad (3)$$

- 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 $i = 2, \ldots, L$.

Algebraic Causal Modeling

- Two widely used frameworks for analyzing causal effects: Causal Diagrams and Potential Outcomes
- Bayesian networks and Causal Diagrams already brought into the realm of Algebraic Statistics [Drton, Sturmfels, and Sullivant (2009), Garcia et. al. (2005), Riccomagno and Smith (2007), and many more]
- Work related to identifiability and latent class models [Drton, Sturmfels, and Sullivant (2009), Fienberg, et. al. (2007), Garcia (2004)]
- Algebraic Flavor of Potential Outcomes
  - Unconfoundedness is basically a statement of conditional independence $\{Y_{i0}, Y_{i1}\} \perp\!\!\!\perp T|X$
  - Consistency is an algebraic condition: $Y = TY_{i1} + (1 - T)Y_{i0}$

Non-identifiable Causal Effects

- Mostly data are observational (or from imperfect experiments):
- Data may also come from different sources
- Is it possible to infer something about ACE?
- Estimating non-identifiable causal effects:
  - Assign a probability measure (prior) to the parameters of latent variables
  - Sample from the posterior distribution consistent with the observed information $\mathbf{T}$
  - Estimate the posterior distribution of Average Causal Effect

Examples

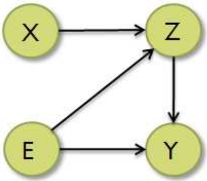

Figure: Violent example from Riccamango and Smith (2007)

X, Z: before & after testosterone levels
E: exposure to a violent movie
Y: arrested for fighting
Exp 1: $P(X)$ and $P(Z|E=1,X)$
Exp 2: $P(Z|Y=1)$, $P(E|Y=1)$ and $P(Y)$

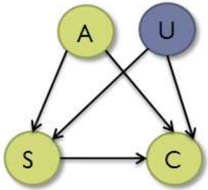

Figure: Speeding and accident

S = speed level and A=age
C = crash
U = unobserved confounder
Obs: $P(C)$, $P(S|C=1)$, $P(A)$, $P(A|C=1)$

Simulation Example

Simulated data of Y_{i0}, Y_{i1}, S, A, U

U	A	S	(Y_{i0}, Y_{i1})	(0,0)	(0,1)	(1,0)	(1,1)
0	0	0		1	16	6	10
		1		3	4	1	1
	1	0		2	8	12	2
		1		1	6	1	2
1	0	0		7	9	6	3
		1		5	19	8	4
	1	0		4	11	10	2
		1		7	20	6	3

ACE = 0.215

Statistical Model - Sensitivity Analysis

- Unspecified Domain of U can be difficult to deal with
- Replace U by a coarsest confounder R_y (Balke and Pearl, 1998, Rubin, and many others)
- For each level of A, R_y has four states, based on the pattern of joint distribution of Potential Outcomes

Y_{i0}	Y_{i1}	R_y
0	0	0
0	1	1
1	0	2
1	1	3

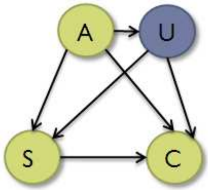

Figure: Causal Model

$R_y = 0$, immune
$R_y = 1$, causative
$R_y = 2$, preventive
$R_y = 3$, doomed

Example - Estimating the posterior of ACE

- Dirichlet prior information specified over latent variables, e.g. $P(R_y|S, C, A)$
- $\mathbf{T}$ is the observed information, in this case, the conditional rates $P(S|C = 1)$, $P(S|A)$ and $P(A|C = 1)$ and the marginals $P(C)$ and $P(A)$
- $\mathcal{M}$ is defined by patterns of R_y (structural zeros)
- Using R4ti2, can sample from the posterior of the joint table $\{C, A, S, R_y, \}$
- Results very sensitive to prior as no new R_y data appears

Posterior of ACE - Result

Computations done using R4ti2 and MCMCpack

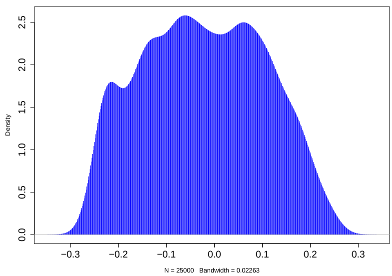

Figure: Non-informative prior

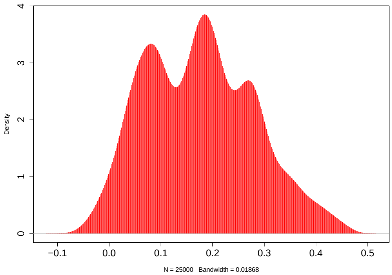

Figure: Informative (skewed prior)

Ecological Inference

- Reconstructing individual behavior from group-level data
- Applications in Political and Social Science, Epidemiology, Geography, Economics,...
- Huge literature of statistical methods starting from Goodman (1953), King (1997), King (2004), Imai, Lu and Strauss (2009)
- Current methods:
  - The Method of Bounds
  - Goodman's Regression
  - King's EI
- Limitations:
  - Work with fractions
  - Almost all methods for 2 by 2 tables
  - Can incorporate only marginal constraints

Inference in voting pattern of different racial groups

X = race $\in \{B, W, H\}$ and Y = voting behavior $\in \{D, R, A\}$
K: number of precincts
$\mathbf{n_k}$: Contingency table associated precinct k.

Partial Information T is a set of linear constraints on each $\mathbf{n_k}$

	Voting			
Race	Demo	Rep	Abstain	Total
Black	?	?	?	n_{1+k}
White	?	?	?	n_{2+k}
Hispanic	?	?	?	n_{3+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	N_k

	Voting		
Race	Demo	Rep	Total
White	$p_{1\|1}$	$p_{2\|1}$	1
Other	$p_{1\|2}$	$p_{1\|2}$	1

- Observed marginals: $\mathbf{T} = \{n_{+i}, n_{j+}\}$
- Observed marginals for K and conditionals over collapsed table for a set $S \subset K$

Several Posterior quantities of interest: μ, Σ, $\sum_k f(\mathbf{n}_k)$, e.g. $\lambda_{ij} = \frac{\sum_k n_{ijk}}{\sum_k n_{i+k} - n_{i3}}$

Bounding Causal Effects

Bound:

$$ACE = q_{20} + q_{21} - q_{10} - q_{11}$$

Subject to:

$$\begin{aligned} p_{1|1}(q_{00} + q_{10}) - p_{0|1}(q_{01} + q_{21}) &= 0 \\ p_{1|0}(q_{20} + q_{30}) - p_{0|0}(q_{11} + q_{31}) &= 0 \\ q_{10}q_{21} - o_1 q_{20} q_{11} & \\ q_{00}q_{31} - o_2 q_{01} q_{30} & \\ \sum_{i=0}^{3}\sum_{j=0}^{1} q_{ij} &= 1 \\ 0 \leq q_{ij} \leq 1 & \end{aligned}$$

Solution using Groebner Basis and Lagrange multipliers:(Using Singular and Maxima)

$$0.0842 \leq ACE \leq 0.5608$$

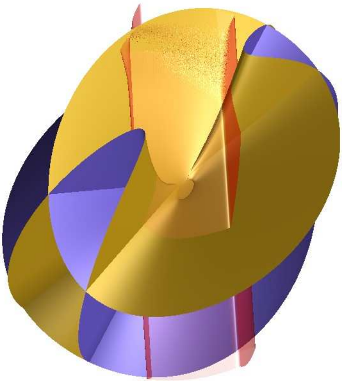

Figure: Surface of ACE

Conclusion

- Make tools from algebraic statistics accessible to applied researchers (R4ti2)
- Framework for inference in non-identifiable models
  - When there is a measured covariate
  - When the structure of the unmeasured confounder is known
  - Sensitivity Analysis for potential confounders
  - Additional assumptions on the structure of counts (in the form of log-linear models)
  - Combine information from disparate sources
  - Ecological Inference
- Privacy and Confidentiality

Future Work

- Issues of data compatibility
- Slow convergence of MCMC algorithm
- Improve sampling $P(\mathbf{p}|\mathbf{n}, \mathbf{T}, \mathcal{M})$
  - Rational parametrization of conditional independence ideal
  - Seems to work for small problems
- A complete 4ti2 (and Singular?) interface for R

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*

Drton, M., Sturmfels, B. and Sullivant, S. (2009).
Lectures on algebraic statistics

Fienberg, S. E., Hersh, P., Rinaldo, A., and Zhou, Y. (2007).
Maximum likelihood estimation in latent class models for contingency table data

Garcia, L.D., Stillman, M., and Sturmfels B. (2005).
Algebraic geometry of Bayesian networks

King, G. (1997).
A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data

Riccomagno, E. and Smith, J. Q. (2007).
Algebraic causality: Bayes nets and beyond

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

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

"Algebraic Statistics is both cool and useful" Bernd Sturmfels

Thank you.