Introduction to Reconstructability Analysis

Martin Zwick
Portland State University, zwick@pdx.edu
Introduction to Reconstructability Analysis

Martin Zwick

Professor of Systems Science

zwick@pdx.edu
http://www.pdx.edu/sysc/research_dmm.html

ISSS 2018, Corvallis, July 22-27
WHAT IS RA?

• Reconstructability Analysis (RA) = a probabilistic graphical modeling methodology

• RA = Info theory + Graph theory

• Graphs, applied to data, are models:
  • node = variable; link = relationship

• RA uses not only graphs (a link joins 2 nodes), but hypergraphs (a link can join >2 nodes)
WHY RA MIGHT BE OF INTEREST TO YOU 1/2

- Can detect many-variable or non-linear interactions not hypothesized in advance, i.e., it is explicitly designed for exploratory search
- Transparent (not black box), easily interpretable
- Designed for nominal variables
- Can also analyze continuous variables via binning
- Prediction/classification, clustering/network models
- Time series, spatial analyses
- Overlaps common statistical & machine-learning methods (but has unique features)
WHY RA MIGHT BE OF INTEREST TO YOU 2/2

• Web-accessible user-friendly software (OCCAM)
• Analyses at 3 levels of refinement:
  – coarse (very fast, many variables)
  – fine (slower, 100s of variables)
  – ultra-fine (slow, < 10 variables)
• Standard application: frequency data \( f(A_i, B_j, C_k, Z_l) \)
• Variety of non-standard capabilities
  – Data: set-theoretic relations & mappings
  – Predict continuous variables
  – Integrate multiple inconsistent data sets
  – Regression-like Fourier version
PAST/PRESENT RA APPLICATIONS

• **BIOMEDICAL**
  Gene-disease association, disease risk factors, gene expression, health care use & outcomes, **dementia**, diabetes, heart disease, prostate cancer, brain injury, primate health, surgery

• **FINANCE-ECONOMICS-BUSINESS**
  Stock market, bank loans, credit decisions, apparel analyses, market segmentation

• **SOCIAL-POLITICAL-ENVIRONMENTAL**
  Socio-ecological interactions, wars, urban water use, rainfall, forest attributes

• **MATH-ENGINEERING**
  Logic circuits, automata dynamics, genetic algorithm & neural network preprocessing, chip manufacturing, pattern recognition, decision analysis

• **OTHER**
  Textual analysis, language analysis
OVERLAP with STATISTICAL, MACHINE LEARNING METHODS

Relation to log linear (LL) (& logistic regression) models & to Bayesian networks (BN)
Where methods overlap, they are equivalent
1. input **data** to RA
   - form of data (cases X variables)
   - data cases indexed by **individual, time, space**
2. model output from RA
3. basics of RA
4. for more information
FORM OF DATA

Variables

• Type: nominal; bin if continuous (continuous DV needn’t be binned)

• Number: few variables to 100s (in principle, to 1000s or more)

• Distinctions:

  directed system
  – Predict/classify a DV (output) from IVs (inputs)

  neutral system
  – No IV-DV distinction: association, clustering / network
FORM OF DATA

- frequency\((A_i, B_j, C_k, Z_l)\) or individual cases

\[
\begin{array}{cccc|c}
A_0 & B_0 & C_0 & Z_0 & 13 \\
A_0 & B_0 & C_0 & Z_1 & 2 \\
A_0 & B_0 & C_1 & Z_0 & 9 \\
A_0 & B_0 & C_1 & Z_1 & 11 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
A_0 & B_0 & C_0 & Z_0 & N \\
\end{array}
\]

\[
\begin{array}{cccc}
A & B & C & Z \\
\hline
\text{case}_1 & A_0 & B_0 & C_0 & Z_0 \\
\text{case}_2 & A_1 & B_2 & C_3 & Z_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{case}_N & A_0 & B_0 & C_0 & Z_0 \\
\end{array}
\]

\(N = \text{sample size}\)

Cases are indexed by individual (in a population), time, or space

\[
frequency(ABCZ) / N = p_{\text{data}}(ABCZ)
\]
### DATA CASES INDEXED BY INDIVIDUAL (#ID)

<table>
<thead>
<tr>
<th>ID</th>
<th>0,0,0,ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>2,1,Ap</td>
</tr>
<tr>
<td>Gender</td>
<td>2,1,Sx</td>
</tr>
<tr>
<td>Education</td>
<td>3,1,Ed</td>
</tr>
<tr>
<td>AgeLastExam</td>
<td>3,1,Ag</td>
</tr>
<tr>
<td>rs1801133</td>
<td>3,1,A</td>
</tr>
<tr>
<td>rs3818361</td>
<td>4,1,B</td>
</tr>
<tr>
<td>rs7561528</td>
<td>3,1,C</td>
</tr>
<tr>
<td>rs744373</td>
<td>3,1,D</td>
</tr>
<tr>
<td>rs6943822</td>
<td>3,1,E</td>
</tr>
<tr>
<td>rs4298437</td>
<td>3,1,F</td>
</tr>
<tr>
<td>rs7012010</td>
<td>3,1,G</td>
</tr>
<tr>
<td>rs11136000</td>
<td>3,1,H</td>
</tr>
<tr>
<td>rs10786998</td>
<td>4,1,J</td>
</tr>
<tr>
<td>rs11193130</td>
<td>4,1,K</td>
</tr>
<tr>
<td>rs610932</td>
<td>3,1,L</td>
</tr>
<tr>
<td>rs3851179</td>
<td>3,1,M</td>
</tr>
<tr>
<td>rs3764650</td>
<td>4,1,N</td>
</tr>
<tr>
<td>rs3865444</td>
<td>4,1,P</td>
</tr>
<tr>
<td>Dementia</td>
<td>2,2,Z</td>
</tr>
</tbody>
</table>

### DEMENTIA EXAMPLE

Z = 0 no disease; Z = 1 disease

<table>
<thead>
<tr>
<th>#ID</th>
<th>Ap</th>
<th>Sx</th>
<th>Ed</th>
<th>Ag</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>112</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>118</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>122</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>123</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
**DATA CASES INDEXED BY TIME**

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-4</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t-3</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>t-2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>t-1</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>t</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-4</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t-3</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>t-2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>t-1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>t</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Values are labels for variable states at particular times  
XYZ = generating variables  
Apply **mask** (here # lags = 2) to data  
Mask adds lagged variables, ABC(t) = XYZ(t-1)  
E.g., A(t-1) = X(t-2), labeled 3

Masking: time series → **atemporal** sample
DATA CASES INDEXED BY SPACE: 1 generating variable

Moore neighborhood

E = DV
A, B, C, D, F, G, H, I = IVs

IVs & DV have 14 possible states

<table>
<thead>
<tr>
<th>#A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>71</td>
<td>71</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>71</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>71</td>
<td>71</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>95</td>
<td>71</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>95</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>90</td>
</tr>
</tbody>
</table>

...
1. input data to RA

2. model output from RA
   
   \( \text{model} = \text{structure (hypergraph)} \text{ applied to data (GT)} \)
   
   \( \text{types of structures (GT)} \)
   
   \( \text{selecting a model (IT)} \)
   
   \( \text{model} = \text{(conditional) probability distribution (IT)} \)
   
3. basics of RA

4. for more information
**MODEL = STRUCTURE APPLIED TO DATA**
A structure (graph or hypergraph) is a set of relationships (GT)

Specific structure AB:BC  General structure

LATTICE OF SPECIFIC STRUCTURES (3 variables)

<table>
<thead>
<tr>
<th>Neutral</th>
<th>df #</th>
<th>Directed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC*</td>
<td>7</td>
<td>ABZ*</td>
</tr>
<tr>
<td>AB:AC:BC</td>
<td>6</td>
<td>AB:AZ:BZ</td>
</tr>
<tr>
<td>AB:AC</td>
<td>5</td>
<td>AB:AZ</td>
</tr>
<tr>
<td>AB:C</td>
<td>4</td>
<td>AB:Z*</td>
</tr>
<tr>
<td>A:B:C*</td>
<td>3</td>
<td>loop</td>
</tr>
</tbody>
</table>

* Reference model is data or independence

# df (degrees of freedom) values are for binary variables
STRUCTURES 4 variables (GT)


**STRUCTURES (GT)**

Combinatorial explosion

<table>
<thead>
<tr>
<th># variables</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># general structures</td>
<td>5</td>
<td>20</td>
<td>180</td>
<td>16,143</td>
</tr>
<tr>
<td># specific structures (where 1 variable is DV)</td>
<td>9</td>
<td>114</td>
<td>6,894</td>
<td>7,785,062</td>
</tr>
<tr>
<td>(1 DV, no loops)</td>
<td>5</td>
<td>19</td>
<td>167</td>
<td>7,580</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

NEED INTELLIGENT HEURISTICS TO SEARCH LATTICE

Can analyze 100s of variables, & for simple models, many more.
TYPES OF STRUCTURES (GT)

FOR PREDICTION / CLASSIFICATION (directed system)

• Variable-based
  – no loops  
    \( IV:ACZ \)  
    many variables (fast) \([coarse]\)  
    simple prediction, feature selection
  – with loops  
    \( IV:ABZ:BCZ \)  
    up to 100s of variables (slow) \([fine]\)  
    better prediction

• State-based  
  \( IV:Z: A_1B_1Z : B_2C_3Z_1 \)  
  < 10 variables (very slow); \([ultra-fine]\)  
  best prediction; detailed models

“IV” = ABC (all IVs); Z = DV
All directed system models include an IV component
TYPES OF STRUCTURES (GT)

Complexity
(degrees of freedom)

Variable-based
No loops With loops

State-based

COARSE FINE ULTRA-FINE
SEARCHING LATTICE OF STRUCTURES

beam search, levels = 3, width = 4  (node = model)
(there are many other search algorithms)
MODEL = PROBABILITY DISTRIBUTION (IT)

for directed system: *conditional* distribution
for neutral system: *joint* distribution
gotten by applying data to a structure

Directed system:

- Model = *calculated* *conditional* probability distribution, e.g., $p_{iv:AZ:BZ}(Z_l | A_i B_j C_k)$

- Distribution gives *rule to predict* DV (Z) from IVs (A,B,C) (e.g., rule = 0 means predict $Z_0$)
SELECTING A MODEL (IT)

1. High information (or low error) in model

   For directed system
   - Info-theory measure: high $\Delta H$, reduction of uncertainty of DV
   - Generic measure: high %correct, accuracy of prediction

2. Low complexity: df, degrees of freedom

3. Information $\leftrightarrow$ complexity tradeoff
   - Statistical significance (Chi-square p-values)
   - Integrated measures: AIC, BIC
     (Akaike & Bayesian Information Criteria)
   - BIC a conservative selection criterion
UNCERTAINTY REDUCTION: SIMPLER EXAMPLE

2 variables: IV=A; DV = Z; T(A:Z)=mutual information (association)

• Uncertainty reduction is like variance explained
  Model AZ = predict Z, i.e., reduce H(Z), by knowing A

• Uncertainty reduced = T(A:Z); uncertainty remaining = H(Z|A)
  \[ \Delta H = \frac{T(A:Z)}{H(Z)} \text{ fractional uncertainty reduction} \] (will express in %)
UNCERTAINTY REDUCTION: SIMPLE EXAMPLE

- $p(Z_1)/p(Z_0) = 1:1$, not knowing $A \rightarrow 2:1$ or 1:2, knowing $A$

- $\Delta H(Z) = T(A:Z) / H(Z) = 8\%$

- 8% reduction in uncertainty is large (unlike variance!)
## SELECTING A MODEL  
**DEMENTIA EXAMPLE**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>model</th>
<th>$\Delta H(%)$</th>
<th>$\Delta df$</th>
<th>%c</th>
<th>$\Delta BIC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable-based (with loops)</td>
<td>BIC</td>
<td>$\text{IV: Ap Z : Ed Z : K Z}$</td>
<td>16</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>State-based</td>
<td>BIC</td>
<td>(model below; each interaction = 1 df)</td>
<td>20</td>
<td>6</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>$\text{IV: Ap}_1 Z : \text{Ed}_0 Z : K_2 Z : \text{Ap}_0 \text{Ed}_2 C_2 Z : \text{Ap}_0 \text{Ed}_1 C_2 K_1 Z : \text{Ap}_0 \text{Ed}_1 C_0 K_1 Z$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Models integrate multiple predicting interactions**

$\text{IV} = \text{ApEdCKL}...$ (all the independent variables);  
\%c( $\text{IV:Z}$ ) = 52
## PROBABILITY DISTRIBUTION DEMENTIA EXAMPLE

| IV | obs p(Z | IV) | calc p(Z | IV) | (p-value) | correct | (p-value) |
|----|---------|-----------|------------|----------|----------|
| Ap | Ed      | K | freq | Z₀ | Z₁ | rule | p rule | # | % | P₀ | ² |
| 0  | 0       | 0 | 4    | 0.0 | 1.000 | .122 | .878 | 1 | 0.131 | 4 | 100.0 | 0.028 |
| 0  | 0       | 1 | 8    | .125 | .875 | .124 | .876 | 1 | 0.033 | 7 | 87.5 | 0.002 |
| 0  | 0       | 2 | 4    | .250 | .750 | .294 | .706 | 1 | 0.409 | 3 | 75.0 | 0.138 |
| 0  | 1       | 0 | 31   | .645 | .355 | .616 | .384 | 0 | 0.198 | 20 | 64.5 | 0.707 |
| 0  | 1       | 1 | 37   | .622 | .378 | .619 | .381 | 0 | 0.147 | 23 | 62.2 | 0.714 |
| 0  | 1       | 2 | 23   | .783 | .217 | .827 | .173 | 0 | 0.002 | 18 | 78.3 | 0.072 |
| 0  | 2       | 0 | 66   | .636 | .364 | .640 | .360 | 0 | 0.023 | 42 | 63.6 | 0.894 |
| 0  | 2       | 1 | 61   | .656 | .344 | .644 | .357 | 0 | 0.025 | 40 | 65.6 | 0.942 |
| 0  | 2       | 2 | 33   | .848 | .152 | .842 | .158 | 0 | 0.000 | 28 | 84.8 | 0.020 |
| 0  | --      | -- | 267  | .648 | .352 | .648 | .352 | 0 | 1 100.0 | 0.571 |
| 1  | 0       | 0 | 1    | .000 | 1.000 | .026 | .974 | 1 | 0.343 | 1 | 100.0 | 0.571 |
| 1  | 0       | 1 | 7    | .143 | .857 | .026 | .974 | 1 | 0.012 | 6 | 85.7 | 0.134 |
| 1  | 0       | 2 | 2    | .000 | 1.000 | .074 | .926 | 1 | 0.228 | 2 | 100.0 | 0.514 |
| 1  | 1       | 0 | 13   | .308 | .692 | .234 | .766 | 1 | 0.055 | 9 | 69.2 | 0.709 |
| 1  | 1       | 1 | 24   | .167 | .833 | .237 | .763 | 1 | 0.010 | 20 | 83.3 | 0.633 |
| 1  | 1       | 2 | 11   | .545 | .455 | .478 | .522 | 1 | 0.884 | 5 | 45.5 | 0.146 |
| 1  | 2       | 0 | 32   | .219 | .781 | .254 | .746 | 1 | 0.005 | 25 | 78.1 | 0.732 |
| 1  | 2       | 1 | 39   | .256 | .744 | .256 | .744 | 1 | 0.002 | 29 | 74.4 | 0.735 |
| 1  | 2       | 2 | 17   | .529 | .471 | .504 | .496 | 0 | 0.973 | 9 | 52.9 | 0.040 |
| 1  | --      | -- | 146  | .281 | .719 | .281 | .719 | 1 | 1 100.0 | 0.735 |
| 413 | .518 | .482 | .518 | .482 | 0 | 291 | 70.5 |
**PROBABILITY DISTRIBUTION DEMENTIA EXAMPLE**

**Decision tree** from conditional probability distribution

(Decrease or decrease of risk given by odds ratios.)

- **Ed₀K₀, Ed₀K₁**: Increased risk of disease; predict $Z₁$
- **Ed₂K₂**: Decreased risk of disease; predict $Z₀$
- **Other EdK**: Predict $Z₀$

- **Ap₀**:
  - **Other EdK**: Predict $Z₀$
- **Ap₁**:
  - **Other EdK**: Predict $Z₁$
  - **Ed₂K₂**: Decreased risk of disease; predict $Z₀$
1. input data to RA
2. model output from RA

3. basic RA algorithms (IT, *inside the black box*)
   - generate model
   - evaluate model

4. for more information
**GENERATE MODEL**

*frequencies shown, not probabilities*

**data:** observed ABC (df=7)

<table>
<thead>
<tr>
<th></th>
<th>C₀</th>
<th>C₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>143</td>
<td>77</td>
</tr>
<tr>
<td>A₁</td>
<td>227</td>
<td>46</td>
</tr>
</tbody>
</table>

**model:** calculated ABC_{AB:BC}

<table>
<thead>
<tr>
<th></th>
<th>C₀</th>
<th>C₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>142</td>
<td>72</td>
</tr>
<tr>
<td>A₁</td>
<td>227</td>
<td>52</td>
</tr>
</tbody>
</table>

**model:** AB:BC (df=5)

<table>
<thead>
<tr>
<th></th>
<th>B₀</th>
<th>B₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>396</td>
<td>259</td>
</tr>
<tr>
<td>A₁</td>
<td>638</td>
<td>185</td>
</tr>
</tbody>
</table>

**1. Projection**

1478

**2. Composition**

**3. Evaluation**

frequencies shown, not probabilities
**GENERATE MODEL**

- **Projection** = sum frequencies or probabilities

- **Composition**

  **Maximize** model entropy *subject to* model constraints

  Model entropy: \[ H(p_{model}) = - \sum p_{model} \log_2 p_{model} \]

  E.g., for model AB:BC, **maximize** \( H(p_{AB:BC}) \) *subject to*

  \[ p_{AB:BC}(AB) = p_{data}(AB) \]
  \[ p_{AB:BC}(BC) = p_{data}(BC) \]

  Composition is **critical computational step**; done

  (a) **Algebraically** (very fast)    loopless models

  (b) **Iteratively** (Iterative Proportional Fitting)    models with loops
**EVALUATE MODEL** (1/2)

- **Evaluation** (1 = data dependent; 2 = data independent)

1. **[ref=data]**
   
   \[ T_{\text{model}} = H_{\text{model}} - H_{\text{data}} = \sum p_{\text{data}} \log_2 \left( \frac{p_{\text{data}}}{p_{\text{model}}} \right) \]

   **[ref=independence]**

   \[ I_{\text{model}} = H_{\text{ind}} - H_{\text{model}} = \sum p_{\text{data}} \log_2 \left( \frac{p_{\text{model}}}{p_{\text{ind}}} \right) \]

   **uncertainty reduction** \[ = H(DV) - H_{\text{model}}(DV \mid IV) \]

2. **[ref=independence]**

   \[ \text{complexity} = \Delta df = df_{\text{model}} - df_{\text{ind}} \]
EVALUATE MODEL (2/2)

Trade off information (or error) & complexity, define best model criterion, via:

Use likelihood ratio Chi-square, $LR = k N T$

- p-values from $\Delta LR$, $\Delta df$, Chi-square table

Or linear combinations of information & complexity

- $\Delta AIC = \Delta LR + 2 \Delta df$
- $\Delta BIC = \Delta LR + \ln(N) \Delta df$
1. input data to RA
2. model output from RA
3. basic RA algorithms

4. for more information
   – DMM (RA) web page
   – Software: OCCAM
   – MORE INFORMATION ON RA
DMM (RA) WEB PAGE

http://pdx.edu/sysc/research-discrete-multivariate-modeling

Research: Discrete Multivariate Modeling

The methods used are also known in the systems literature as “reconstructability analysis” (RA). RA overlaps significantly with the fields of logic design and machine learning and with log-linear statistical modeling. The papers “Whales and Partis in General Systems Methodology” and “An Overview of Reconstructability Analysis” listed below offer a concise review of RA methodology.

Projects

Theory/Methodology

OCCAM: RA software for data analysis & data mining

Occam2 (web accessible; try it out)

User manual (PDF)

EDA: Extended Dependency Analysis

Heuristic RA search for loopless models

Download executable, sample files, and documentation (for Windows)

RA utility programs
SOFTWARE: OCCAM (access on DMM page)

Occam is a Discrete Multivariate Modeling (DMM) tool based on the methodology of Reconstructability Analysis (RA). Its typical usage is for analysis of problems involving large numbers of discrete variables. Models are developed which consist of one or more components, which are then evaluated for their fit and statistical significance. Occam can search the lattice of all possible models, or can do detailed analysis on a specific model.

In Variable-Based Modeling (VBM), model components are collections of variables. In State-Based Modeling (SBM), components identify one or more specific states or substrates.

Occam provides a web-based interface, which allows uploading a data file, performing analysis, and viewing or downloading results.

- Run Occam
- For basic operation instructions, please see the manual: PDF
- Sample data files. You can download these to local files on your computer, then upload them via the Occam Web interface.
- A Neutral System
- A Directed System
- Links:
  - Dr. Zwick's DMM Research Page
  - Systems Science Graduate Program
  - Occam-users mailing list (discussion)
  - Occam-news mailing list (announcements)
- Contacts:
  - Occam feedback email address
  - Dr. Martin Zwick, Systems Science
  - Joe Fusion, Graduate Assistant, Systems Science
**BASIC OCCAM ACTIONS**

- **Search** = exploratory modeling, examine many models, find best or good ones  
  (OCCAM actions: Search, SB-Search)

- **Fit** = confirmatory modeling, look at one model in detail (see probability distribution) & use for prediction  
  (OCCAM actions: Fit, SB-Fit)

(OCCAM actions: Show Log, Manage Jobs = managerial functions)
INFORMATION ON RA

• Review articles on DMM page
  – “Wholes & Parts in General Systems Methodology” (accessible)
  – “An Overview of Reconstructability Analysis” (encompassing)


• International Journal of General Systems

• Kybernetes, Vol. 33, No. 5/6 2004: special RA issue
• OCCAM is available for use
  (but consult with me before doing anything other than variable-based models without loops)

• Plan to make OCCAM open-source; contact me if you would like to be involved

• zwick@pdx.edu

• Thank you.
**UNCERTAINTY REDUCTION: DEMENTIA EXAMPLE**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>model</th>
<th>$\Delta H(%)$</th>
<th>$\Delta df$</th>
<th>%c</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>IV:ApZ:EdZ:CZ</td>
<td>16</td>
<td>5</td>
<td>70</td>
</tr>
</tbody>
</table>

$$\Delta H = \frac{T_{IV:ApZ:EdZ:CZ}(ApEdC:Z)}{H(Z)} = 14\%$$

![Venn Diagram](https://via.placeholder.com/150)
In 5-generating-variables spatial example, model could be:

IV: ABCD Z:EFG Z:HIJ Z