Exploratory Modeling of TBI Data

Martin Zwick
& S. Kolakowsky-Hayner, N. Carney, M. Balamane, T. Nettleton, D. Wright

Systems Science Program
Portland State University

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

2016 TBI Symposium OHSU Sept 16-17, 2016
• **Data Analytics/Occam Subproject**, Portland State University
  – Martin Zwick, co-PI
  – (Wayne Wakeland, PI of Dynamic Model Initiative)
  – Programmers: Forrest Alexander, Peter Olson

• **Brain Trauma Evidence-Based Consortium (BTEC)**
  • **Stephanie Kolakowsky-Hayner**, Brain Trauma Foundation, BTEC project head
    – Assistant Program Manager: Maya Balamane

• **Nancy Carney**, OHSU, BTEC founder & previous BTEC project head
  • Research assistant: Tracie Nettleton

• Funded by DoD via BTF & Stanford

1. Exploratory modeling with Occam

2. Sample results on Preece, Wright data sets
1. Exploratory modeling with Occam

- Exploratory modeling (data mining) with Reconstructability Analysis (RA):
  - to contribute to a clinically-useful TBI classification system & other BTEC projects
  - to extract additional information from past studies
Rationale for exploratory modeling

• Most studies are confirmatory, testing only specific hypotheses. Since studies are expensive & time-consuming, useful to explore what might be discovered in the data.

• Exploratory studies can find unexpected non-linear & many-variable interaction effects (should then be tested in confirmatory mode).

• Exploratory studies (by data analysts) are unbiased.
Why RA & Occam software

• Explicitly designed for exploratory modeling
  – Analyzes both nominal & continuous (binned) variables
  – Easily interpretable; standard text input; web-accessible, emails results to user; available for research use

• Other statistical & machine-learning methods (log-linear, logistic regression, Bayesian networks, classification trees, support vector machines, neural nets) not well designed for exploration, or have limited model types, or have difficulty with nominal variables or with stochasticity
What RA is

• Reconstructability Analysis (RA) = Information theory + Graph theory, a probabilistic graphical modeling technique

• RA model = a (conditional) probability distribution simpler (fewer df) than the data, capturing much of the information in the data
Approach (1/2)

2 types of model searches

• **Neutral**: find relationships among all variables (‘clustering’)

• **Directed**: predict DVs from IVs (‘classification’); want high
  – **Accuracy** (information captured) measured by
    • $\% \Delta H = \%$ reduction of uncertainty (info measure like variance)
    • $\% c = \%$ correct in prediction (a general measure)
  – **Simplicity** = low $\Delta df$ (trades off with accuracy)
  – Integrate w’ BIC, conservative model-selection criterion
Approach (2/2)

3 degrees of refinement of RA search

Complexity
(degrees of freedom)

Variable-based
No loops
COARSE

With loops
FINE

State-based
ULTRA-FINE
Occam input file (partial, Preece) (note missing data)
2. Sample results

2.1 Preece data: analysis completed
   auto accidents

2.2 Wright (PROTECT) data: analysis underway
   auto/motorcycle/bike accidents, hit pedestrians, falls

Other data sets to follow
2.1 *Preece data*

- **52 variables**

- **Variable types**
  - $P =$ *patient* characteristics (17 variables)
  - $Y =$ *symptoms* (25): subjective reports
  - $G =$ *signs* (4): objective indicators
  - $C =$ *cognitive* deficits (5)
  - $N =$ *neurologic* deficits (1)

- $N = 337$; reduces to 175 or less if exclude missing data
**Directed searches**

- DVs (cognitive, neurological deficit variables)
  - #bins excludes missing values

<table>
<thead>
<tr>
<th>#bins</th>
<th>N</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Cdg</td>
<td>255 <strong>Digit Symbol Substitution neuropsychological test</strong></td>
</tr>
<tr>
<td>6</td>
<td>Cnr</td>
<td>210 <strong>Spatial Reaction Time normalized for age and sex</strong></td>
</tr>
<tr>
<td>6</td>
<td>csr</td>
<td>214 Spatial Reaction Time test: how quickly patient responds to visual stimuli</td>
</tr>
<tr>
<td>3</td>
<td>Nlr</td>
<td>209 <strong>LogMAR  Log of Minimum Angle of Resolution (visual acuity)</strong></td>
</tr>
</tbody>
</table>
### Cnr coarse, fine, ultra-fine searches

#### Predict Cnr: reaction time, normalized by age, sex (rebin |Cnr| = 2: ~ 50-50)

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Δdf</th>
<th>p</th>
<th>%ΔH</th>
<th>%c</th>
<th>N=175</th>
</tr>
</thead>
<tbody>
<tr>
<td>COARSE, single component predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cdg Gpt Cnr</td>
<td>3</td>
<td>0.00</td>
<td>10.6</td>
<td>64.6</td>
<td>BIC, AIC</td>
</tr>
<tr>
<td>Pph Cdg Gpt Cnr</td>
<td>7</td>
<td>0.00</td>
<td>13.1</td>
<td>66.9</td>
<td>IncrP</td>
</tr>
<tr>
<td>Cnr (independence=reference)</td>
<td>0</td>
<td>1.00</td>
<td>0.0</td>
<td><strong>50.9</strong></td>
<td></td>
</tr>
<tr>
<td>FINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cdg Cnr : Gpt Cnr</td>
<td>2</td>
<td>0.00</td>
<td>8.8</td>
<td>64.6</td>
<td>BIC</td>
</tr>
<tr>
<td>Pri Cnr : Pph Cnr : Cdg Gpt Cnr</td>
<td>6</td>
<td>0.00</td>
<td>14.7</td>
<td>70.3</td>
<td>AIC</td>
</tr>
<tr>
<td>Pye Cnr : Pph Cnr : Cdg Gpt Cnr</td>
<td>5</td>
<td>0.00</td>
<td>12.9</td>
<td>67.4</td>
<td>IncrP</td>
</tr>
<tr>
<td>ULTRA-FINE (state-based model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pph₁ Cdg₁ Cnr : Cdg₀ Gpt₁ Cnr</td>
<td>2</td>
<td>0.00</td>
<td>12.4</td>
<td>64.8</td>
<td>BIC</td>
</tr>
<tr>
<td>Cnr (independence=reference)</td>
<td>0</td>
<td>1.00</td>
<td>0.0</td>
<td><strong>50.9</strong></td>
<td></td>
</tr>
</tbody>
</table>
**Cnr ultra-fine (state-based) model**

**Reaction time model:** \[ P_{\text{ph}_1} \cdot C_{\text{dg}_1} \cdot C_{\text{nr}} : C_{\text{dg}_0} \cdot G_{\text{pt}_1} \cdot C_{\text{nr}} \]

Odds (high is good) = \( \frac{C_{\text{nr}_0}}{C_{\text{nr}_1} \text{(model)}} = p(\text{fast, i.e., normal})/p(\text{slow}) \)

Pph\(_1\) previous head injury, Cdg\(_1\) high digit score; Gpt\(_1\) amnesia

<table>
<thead>
<tr>
<th>IV states</th>
<th>N</th>
<th>Cnr(_0)</th>
<th>Cnr(_1)</th>
<th>Cnr(_0)</th>
<th>Cnr(_1)</th>
<th>Odds</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0</td>
<td>20</td>
<td>0.40</td>
<td>0.60</td>
<td>0.52</td>
<td>0.48</td>
<td>1.1</td>
<td>.92</td>
</tr>
<tr>
<td>0 0 1</td>
<td>19</td>
<td>0.16</td>
<td>0.84</td>
<td>0.16</td>
<td>0.84</td>
<td>(0.2)</td>
<td>(.00)</td>
</tr>
<tr>
<td>1 0 0</td>
<td>30</td>
<td>0.57</td>
<td>0.43</td>
<td>0.52</td>
<td>0.48</td>
<td>1.1</td>
<td>.90</td>
</tr>
<tr>
<td>1 0 1</td>
<td>18</td>
<td>0.17</td>
<td>0.83</td>
<td>0.16</td>
<td>0.84</td>
<td>(0.2)</td>
<td>(.00)</td>
</tr>
<tr>
<td>0 1 0</td>
<td>24</td>
<td>0.50</td>
<td>0.50</td>
<td>0.52</td>
<td>0.48</td>
<td>1.1</td>
<td>.91</td>
</tr>
<tr>
<td>0 1 1</td>
<td>13</td>
<td>0.61</td>
<td>0.39</td>
<td>0.52</td>
<td>0.48</td>
<td>1.1</td>
<td>.93</td>
</tr>
<tr>
<td>1 1 0</td>
<td>38</td>
<td>0.76</td>
<td>0.23</td>
<td>0.73</td>
<td>0.27</td>
<td>2.7</td>
<td>.01</td>
</tr>
<tr>
<td>1 1 1</td>
<td>14</td>
<td>0.64</td>
<td>0.36</td>
<td>0.73</td>
<td>0.27</td>
<td>2.7</td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td>176</td>
<td>0.51</td>
<td>0.49</td>
<td>0.51</td>
<td>0.49</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>
Reaction time odds (probability fast/ probability slow) & p-values relative to marginal prob. (odds = 1)

Cnr decision tree from conditional probabilities

Digit symbol score

Previous head injury

Amnesia

low

yes .2 .00

no

1.1 .92

normal

no 1.1 .91

yes 2.7 .01, .09
For low performance on digit symbol test, amnesia predicts slow reaction time.

For normal performance on digit symbol test, previous head injury increases the probability of fast (normal) reaction time. THIS IS ANOMALOUS.

– Need to see if it would be replicated in another data set.
– Possible explanation: prior exposure to Reaction Time test introduces a practice effect.
2.2 *Wright data*

- **560 variables** (302 variables within 1st two weeks)

- **Variable types**
  - $A =$ **admin** (32 variables) #1-32
  - $P =$ **patient characteristics** (134 variables) #405-538
  - $Y =$ **symptoms** (8 variables): subjective reports #551-558
  - $G =$ **signs** (13 variables): objective indicators #539-550, 560
  - $C =$ **cognitive deficits** (6 variables) #33-38
  - $N =$ **neurologic deficits** (367 variables) #39-404, 559

- **N = 882 patients**
Two lines of current exploration (1/2)

• Predict DV = mortality at 2 weeks (N=764)
• No surprises: GCS scores, days 2, 4, 9, are best predictors.
Two lines of current exploration (2/2)

• Look for a possible progesterone effect

• Effects expected but not found in Wright study
• Didn’t systematically look for possible complex effects

• RA detects a possible predictive interaction effect
• Likely an artifact, but under investigation
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 "Wholes and Parts 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

Occam3 (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

Below is the lattice of structures for a 4-variable directed system with 1 dependent variable (output). Boxes = relations; lines = variables; bold lines = the dependent variable.
RA software (Occam)

Occam is a Discrete Multivariate Modeling (DMM) tool based on the methodology of Reconstructibility 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
PSU COURSES

• Discrete Multivariate Modeling (DMM)
  theory course (SySc 551)
  Fall 2016 (1st class: Sept 27)

• Data Mining with Information Theory (DMIT)
  data analysis project course (DMM not a prerequisite)
  Winter 2017
• THANK YOU