Data Mining with Information Theory (DMIT)

SySc 531 Jan 4, 2022

Prof. Martin Zwick zwick@pdx.edu

https://web.pdx.edu/~zwick/

Terminology

- DATA MINING = exploratory modeling (machine learning): finding relationships between variables in data
- INFORMATION THEORY + graph theory = Reconstructability Analysis (RA),
 a.k.a discrete multivariate modeling (DMM)
- OCCAM = RA software for course

Projects

- All projects must use OCCAM
- Projects can either (a) use OCCAM alone, or (b) use OCCAM + another data mining method & compare results. (a) is standard for this course. If you do (b), state-based (SB) OCCAM analysis is not required
- Support of other methods is not part of this course

DMM section of MZ Selected Works: resource for this course

https://works.bepress.com/martin_zwick/



4

More information on RA

- Review articles on SW page & Canvas
 - "Wholes & Parts in General Systems Methodology" https://works.bepress.com/martin_zwick/52/
 - "An Overview of Reconstructability Analysis" https://works.bepress.com/martin_zwick/57/
- Krippendorff, Klaus (1986). Information Theory. Structural Models for Qualitative Data (Quantitative Applications in the Social Sciences Monograph #62). New York: Sage Publications (at PSU Bookstore).
- International Journal of General Systems
- Kybernetes, Vol. 33, No. 5/6 2004: special RA issue

Past/present applications (some on DMM page)

• BIOMEDICAL

Gene-disease association, disease risk factors, gene expression, health care use & outcomes, medical records (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, optimization, neural networks, chip manufacturing, pattern recognition, decision analysis

• OTHER

Textual analysis, language analysis

Relation to other methods

- RA related to log-linear methods, graphical modeling, e.g., Bayesian networks, logistic regression, etc., competitive with NN, other machine learning methods
- RA is explicitly designed for exploratory search & has unique features not available in those methods
- RA is transparent: readily interpretable

Objectives & deliverables

ESSENTIAL GOALS

- To give you concrete experience of data mining
- To teach you to use OCCAM to analyze data
- To analyze your data

DELIVERABLE

• Report & presentation of your analysis

If you're ambitious, consider (this is totally optional!)

- Comparing RA to another method
- Making your report into a conference or journal paper

Course prerequisites

- YOU <u>HAVE DATA</u> TO ANALYZE of a type suitable for OCCAM (rectangular, adequate sample size)
- <u>Some knowledge of probability/statistics</u>
- Discrete Multivariate Modeling (DMM, SySc 551) NOT prerequisite.
- DMM will probably be taught in academic year 2022-2023.
- DMIT presents only theory needed to understand OCCAM output

Software tool: OCCAM: dmit.sysc.pdx.edu

🕙 Occam3 - Mozilla Firefox										
<mark>File Edit View History Bookmarks Tools Help 🔀 -</mark> Google										
Occam3 +										
S dmm.sysc.pdx.edu										
	Contact PSU PSU FAQs									
Portland State	search PSU 📀 💽	Ð								
	SEARCH									
FUTURE STUDENTS CURRENT STUDENTS FACULTY + STAFF ALUMNI + FRIENDS	🛛 I want to 🔽 Ϛ	6								



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 substates.

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

- <u>Run Occam</u>
- For basic operation instructions, please see the manual: <u>PDF</u>
- Sample data files. You can download these to local files on your computer, then upload them via the Occam Web interface.

<u>A Neutral System</u> <u>A Directed System</u>

- Links: Dr. Zwick's <u>DMM Research Page</u> <u>Systems Science Graduate Program</u> <u>Occam-users mailing list (discussion)</u> <u>Occam-news mailing list (announcements)</u>
- Contacts: <u>Occam feedback email address</u> <u>Dr. Martin Zwick, Systems Science</u> <u>Joe Fusion, Graduate Assistant, Systems Science</u>

Give to PSU | PSU FAQs | Contact PSU | Find People | Maps/Directions | PSU Sitemap | © 2000-2012

OCCAM server

- dmit is the OCCAM server for this course; openly accessible over the web
- If class overloads dmit, another server will be made available
- OCCAM is now open source; at <u>https://github.com/occam-ra/occam</u>

OCCAM manual

OCCAM: A Reconstructability Analysis Program (Organizational Complexity Computation and Modeling)	Occam User's Manual 1/19/2021	2			
	Table of Contents				
	I. FOR INFORMATION ON RECONSTRUCTABILITY ANALYSIS	3			
Project Director: Martin Zwick	II. ACCESSING OCCAM	3			
	III. SEARCH INPUT	4			
	IV. SEARCH OUTPUT	15			
	V. STATE-BASED SEARCH	19			
Past Programmers:	VI. FIT INPUT	20			
Heather Alexander, Joe Fusion, Kenneth Willett ¹	VII. FIT OUTPUT	24			
	VIII. STATE-BASED FIT	31			
	IX. SHOW LOG	31			
Systems Science Program	X. MANAGE JOBS	31			
Portland State University Portland OR 97207	XI. FREQUENTLY ASKED QUESTIONS	31			
	XII. ERROR AND WARNING MESSAGES				
This manual was last revised on 19 January 2021.	XIII. KNOWN BUGS & INFELICITIES; LIMITATIONS	37			
	XIV. PLANNED BUT NOT-YET-IMPLEMENTED FEATURES	38			
Occam version 3.4.1, copyright 2000-2020.	APPENDIX 1. REBINNING (RECODING)	40			
	APPENDIX 2. MISSING VALUES IN THE DATA	42			
	APPENDIX 3. ADDITIONAL PARAMETERS IN THE INPUT FILE	42			
	APPENDIX 4. ZIPPING THE INPUT FILE	43			
	APPENDIX 5. COMPARE MODE	43			
	APPENDIX 6. CACHED DATA MODE	45			

 $^{^1}$ Ken Willett totally rewrote earlier versions of Occam. His version was originally called "Occam3" to distinguish it from these earlier Occam incarnations; the "3" has finally been dropped in this manual

OCCAM initial screen

Ēi	le <u>E</u> dit <u>V</u> iew Hi <u>s</u> tory <u>B</u> oo	kmarks <u>T</u> ools <u>H</u> elp 🔍 S	earch		
	Occam 3	× +			
é	🖬 🔳 🌾 🛈 dmit.s	ysc. pdx.edu /weboccam.cgi			순 自 俞 ∢ 역
6	Portla	and State	P _Y		
	Occam	version 3.3.11 — M	lon Dec 19 20:59:01 2016		
	© Do Search	© Do SB-Search	🛇 Do Fit	🔘 Do SB-Fit	🔘 Do Compare
	🔿 Show Log	🔘 Manage Jobs			🔲 Cached Data Mode
	© 2000-2016				

OCCAM actions

- Search = exploratory modeling, examine many models, find best or good ones (OCCAM actions Search, SB-Search)
- Fit = look at <u>one</u> model in detail & use for prediction; confirmatory modeling (OCCAM actions Fit, SB-Fit)

Compare: specific to one PhD project

Show log, Manage jobs: managerial functions

RA model types (1/3)

- Models are
 - Variable-based (VB), or
 - State-based (SB)

- A VB or SB model
 - Is loopless, or
 - Has loops

RA model types (e.g., 4 variables: A,B,C,Z) (2/3)

- Variable-based
 - loopless ABC:ABZ

- *many* variables *COARSE* simple prediction, *feature selection* FAST
- with loops ABC:AZ:BZ
- up to 100s variables *FINE* better prediction, *detailed analysis* SLOW (exercise caution in runs)
- State-based
 - -- usually W' loops
 - ABC:Z: $A_1Z:B_2Z_1$

- < 10 variables ULTRA-FINE
- S best (most precise) prediction
 V. SLOW (exercise super caution)

RA model types (3/3)



What a model is

- Some models:

 ABC:ABZ
 VB loopless
 ABC:AZ:BZ
 VB with loops
 ABC:Z: A₁Z:B₂Z₁
 SB (usually with loops)
- A model = set of relations, separated by colon ":"
- Relations are predictive (ABZ) or not (ABC = "IV")
- Multiple predictive relations must be integrated
- A relation is a joint probability distribution

Lattice of models (e.g., 4 variables: A,B,C,Z)

- Top model = data = "saturated model" = most complex, e.g., ABCZ
- Bottom model = "independence model" = least complex
 - With IV | DV distinction (directed): ABC:Z
 - Without IV | DV distinction (neutral) : A:B:C: ...
 - Or, bottom could be uniform distribution
- Search up from bottom, or down from top,
 - Or up/down from some other model
- Starting model (& search direction) VS. reference model

OCCAM search (VB, SB) input page

Portland State

OCCAM version 3.3.11 - Fri Jan 1 15:50:55 2016										
◯ Do Fit	SB-Fit O Do O Do Compare O Show Log Jobs									
General Settings										
Data File: Browse No file selected.										
Starting Model:										
Composition Method:	Standard O Back Projection (Fourier)									
Reference Model:	Default Top Bottom Starting Model									
Search Settings										
Models to Consider:	○ All									
Search Direction:	O Default ○ Up ○ Down O Down									
During Search, Sort By:	BIC ○ AIC ○ Information ○ Alpha ○ % Correct									
When Searching, Prefer:	Larger Values Smaller Values									
Search Width: (models to keep at each level)	3									
Search Levels:	7 (leave blank to use settings from data file)									
	Report Settings									
In Report, Sort By:	Information ○ Alpha ○ dDF ○ Level ○ % Correct ○ BIC ○ AIC AIC AIC									
In Report, Sort:	Descending O Ascending									
Include in Report:	☑ H ☑ dLR ☑ Alpha ☑ % dH(DV) ☑ dAIC ☑ dBIC									
	BP-based Transmission 🗵 Incremental Alpha									
	☑ % Correct ☑ % Coverage of Data ☑ % Missing in Test									
Return data in spreadsheet format										
Print option settings 🛛 🖾 (but don't pr	int variable definitions)									
Use inverse notation for models										
Run in Background, Email Results To:										
Subject line for email (optional):	Send									

20

Variables & cases

- <u>Variables</u>: so far up to about 500 IVs Need <u>at least 5</u>
 Best is bigger, but not too big (>10, <100)
 IV/DV (input/output) distinction: directed
 No distinction: neutral This course: directed only
- Will usually (not always) name the DV "Z" (or "X")
- <u>Cases</u> (sample size): so far up to 6.5 x 10⁶ Need <u>at least 100</u> Best is bigger, but not too big (>1K, <<1M)

Data (e.g., IVs = A,B,C; DV = Z)

Contingency table or
 frequency (A_i, B_j, C_k, Z_l)

				frequency
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
•••	•••	•••		•••

No frequency, just

cases X variables

	Α	В	С	Ζ
case ₁	A ₀	B_0	C_0	Z_0
case ₂	A_1	B ₂	C ₃	Z_1
•••				
case _N	A_0	B_0	C_0	Z_0

- <u>Cases</u> = individuals, instances, time or space values
- <u>Variables</u> = IVs & DV are nominal

If continuous, bin (e.g., with Excel binning program) (In *variants* of standard RA, DV can be continuous)

Sample input file: weisdorf.txt) (for demo#1)

- ·:action search
- :nominal season, 2,1,a absentee, 2,0,b sqfeet, 8,1,c valued, 12,1,d Rectangular Snip yrbuilt, 12,1,e zip, 40,0,f ccf, 16,0,w normccf, 16,2,x
- :no-frequency

:da	ta						
0	0	1	2	9	39	1	1
0	1	2	3	9	19	0	0
0	1	1	3	5	19	2	2
0	1	1	4	7	19	0	0
0	0	1	2	0	1	3	2
0	1	1	1	0	1	3	2
0	1	1	3	9	39	1	1
0	1	3	5	7	19	1	0
0	1	1	2	1	39	0	0
0	0	0	1	5	39	0	0
0	1	1	1	1	39	1	1
0	1	2	3	6	19	1	1
0	0	2	2	0	1	1	1
0	1	2	4	5	19	2	2
0	1	2	3	1	39	2	2
0	1	2	3	6	19	2	2
0	1	0	1	3	39	1	0
0	1	1	3	5	39	1	1
0	1	0	2	0	39	1	1
0	1	1	2	1	39	0	0
0	1	1	4	3	19	2	2

Preparing an OCCAM input file

- Include all variables you have (you can always tell OCCAM to ignore some); make DV the *last* variable
- In binning variables (by yourself or with utility program), 3 bins is a reasonable default, but it's better to use more bins (e.g., 12) since you can rebin (aggregate bins) on the fly in the OCCAM input file; see Appendix of OCCAM manual on rebinning
- Code missing values as "."
- Use comments "#" in input file

OCCAM search output (1/4)

 Search many models: select best w ('width') models at each of I 'levels'

• Output all selected w*l models

• OCCAM summarizes the best of these

OCCAM search output (2/4)



Upwards search, increasing complexity, width=4, levels=3

OCCAM search output (3/4)

- OCCAM summarizes:
- Of width*levels models found in search, it indicates the best three by different criteria:
 - (i) highest ΔBIC
 - (ii) highest ΔAIC
 - (iii) highest information with OK p-values (both cumulative & incremental)

Search Output for weisdorf.txt VB models without loops (search sort on bic)(demo#1)

prace phace pize	22/25
Sample Size	158018
H(data)	9.61708
H(IV)	7.46711
H (DV)	2.35084
T(IV:DV)	0.200864
IVs in use (4)	ACDE
DV	Х

Searching levels:

1 : 4 new models, 3 kept; 5 total models, 4 total kept; 0 kb memory used; 0.0 seconds, 0.0 total 2 : 6 new models, 3 kept; 11 total models, 7 total kept; 828 kb memory used; 0.1 seconds, 0.1 total 3 : 3 new models, 3 kept; 14 total models, 10 total kept; 2096 kb memory used; 0.1 seconds, 0.1 total 4 : 1 new models, 1 kept; 15 total models, 11 total kept; 2592 kb memory used; 0.1 seconds, 0.2 total

5 : 0 new models, 0 kept; 15 total models, 11 total kept; 2592 kb memory used; 0.0 seconds, 0.2 total

ID	MODEL	Level	H	dDF	dLR	Alpha	Inf	8dH(DV)	dAIC	dBIC	Inc.Alpha	Prog.	8C (Data)	<pre>%cover</pre>
11	ACDEX	4	9.6171	31665	44001.1153	0.0000	1.00000000	8.5444	-19328.8847	-335043.6345	1.0000	8	39.7733	52.9356
10*	IV:ACEX	3	9.6613	2865	34314.0074	0.0000	0.77984404	6.6633	28584.0074	18.6274	0.0000	5	39.1474	98.9583
9*	IV:ACDX	3	9.6654	2625	33417.9888	0.0000	0.75948049	6.4893	28167.9888	1995.5202	0.0000	6	39.1848	77.2727
8*	IV:ADEX	3	9.6676	3945	32943.0006	0.0000	0.74868558	6.3970	25053.0006	-14280.4808	0.0000	5	39.2012	100.0000
7*	IV:ACX	2	9.6785	225	30539.7263	0.0000	0.69406710	5.9304	30089.7263	27846.3718	0.0000	3	38.9595	100.0000
6*	IV:ADX	2	9.6921	315	27574.9068	0.0000	0.62668654	5.3546	26944.9068	23804.2106	0.0000	2	38.9013	100.0000
5*	IV:AEX	2	9.7456	345	15841.4381	0.0000	0.36002356	3.0762	15151.4381	11711.6279	0.0000	3	38.7437	100.0000
4*	IV:CX	1	9.7471	105	15514.9322	0.0000	0.35260316	3.0128	15304.9322	14258.0335	0.0000	1	38.7753	100.0000
3*	IV:AX	1	9.7606	15	12561.7145	0.0000	0.28548628	2.4393	12531.7145	12382.1576	0.0000	1	38.7437	100.0000
2*	IV:DX	1	9.7617	150	12309.9942	0.0000	0.27976550	2.3904	12009.9942	10514.4246	0.0000	1	38.8070	100.0000
1*	IV:X	0	9.8179	0	0.0000	1.0000	0.00000000	0.0000	0.0000	0.0000	0.0000	0	38.7437	100.0000
ID	MODEL	Level	Н	dDF	dLR	Alpha	Inf	8dH(DV)	dAIC	dBIC	Inc.Alpha	Prog.	8C (Data)	<pre>%cover</pre>

Best	Model(s)	bу	dBIC:											
7*	IV:ACX	2	9.6785	225	30539.7263	0.0000	0.69406710	5.9304	30089.7263	27846.3718	0.0000	3	38.9595	100.0000
Best	Model(s)	by	dAIC:											
7*	IV:ACX	2	9.6785	225	30539.7263	0.0000	0.69406710	5.9304	30089.7263	27846.3718	0.0000	3	38.9595	100.0000
Best	Best Model(s) by Information, with all Inc. Alpha < 0.05:													
10*	IV:ACEX	3	9.6613	2865	34314.0074	0.0000	0.77984404	6.6633	28584.0074	18.6274	0.0000	5	39.1474	98.9583

Sample input file: dementia05.txt (for demo#2,#3)

#Fifth test set of OHSU data 20 June 2012
#Corrected transform of variable I rs12248379
#Corrected transform of APOE
#Updated imputed data
#Deleted var I and dropped records with no var C data
#Added stats on missing. Blank = none missing

:nominal				
ID	,0,0,ID			
APOE	,2,1,AP			
Gender	,2,1,5X			
Education	,3,1,ED			
AgeLastExam	,3,1,AG			
rš1801133	,3,1,A			
rs3818361	,4,1,B	#	missing	3
rs7561528	,3,1,C		-	
rs744373	,3,1,D			
rs6943822	,3,1,E			
rs4298437	,3,1,F			
rs7012010	,3,1,G			
rs11136000	,3,1,Н			
rs10786998	,4,1,]	#	missing	9
rs11193130	,4,1,К	#	missing	11
rs610932	,3,1,L			
rs3851179	,3,1,M			
rs3764650	,4,1,N	#	missing	2
rs3865444	,4,1,P	#	missing	9
CaseControl	,2,2,Z			

:no-frequency

:data													
#IndID	APOE	GENDER	EDU	ALE	A	В	C	D	E	F	G	н	J
101	0	0	2	2	1	1	0	1	2	2	1	1	2
103	0	0	2	1	0	2	2	0	1	1	1	2	2
111	0	1	2	1	2	2	1	1	0	1	1	2	1
112	0	0	2	2	2	2	1	1	1	2	1	1	0
118	0	1	0	2	2	2	2	0	0	1	1	1	
120	0	1	2	2	1	2	1	1	0	1	1	2	1
121	0	0	2	2	2	2	1	1	2	0	0	0	2
122	0	0	1	2	1	2	1	1	2	0	0	2	2
123	0	0	2	2	2	2	2	0	1	1	0	0	2
126	0	0	2	2	2	2	2	0	1	1	2	2	2
127	0	1	2	2	0	2	1	1	1	1	1	2	2
128	0	0	2	2	1	2	1	1	0	1	1	2	
129	1	1	2	2	0	2	1	0	1	0	1	2	0
132	0	0	2	2	2	2	1	0	0	1	1	2	1
134	0	1	2	2	0	0	2	0	0	1	Ľ1	1	0
135	0	0	2	2	2	1	1	0	0	2	1	1	• 2

Search Output for dementia05.txt VB models with loops (search sort on info)(demo#2)

ID	MODEL	Level	H	dDF	dLR	Alpha	Inf	8dH (DV)	dAIC	dBIC	Inc.Alpha	Prog.	<pre>%C (Data)</pre>	<pre>%cover</pre>
22	IV:ApZ:EdCZ:EdKZ:LZ	7	9.5221	20	120.2026	0.0000	0.20476589	20.4766	80.2026	-0.7921	0.0505	19	72.4057	52.7778
21	IV:ApZ:EdCZ:EdKZ:BZ	7	9.5224	21	120.0402	0.0000	0.20448919	20.4489	78.0402	-7.0043	0.1093	18	72.6415	38.5417
20	IV:ApZ:EdKZ:EdLZ:CZ	7	9.5224	20	120.0236	0.0000	0.20446093	20.4461	80.0236	-0.9711	0.0890	17	71.9340	52.7778
19	IV:ApZ:EdCZ:EdKZ	6	9.5323	18	114.2291	0.0000	0.19459005	19.4590	78.2291	5.3339	0.1469	16	71.6981	76.3889
18	IV:ApZ:EdKZ:BZ:CZ	6	9.5352	17	112.5147	0.0000	0.19166949	19.1669	78.5147	9.6692	0.1656	16	72.1698	38.5417
17	IV:ApZ:EdKZ:CZ:LZ	6	9.5361	16	111.9806	0.0000	0.19075974	19.0760	79.9806	15.1849	0.0053	14	72.4057	52.7778
16*	IV:ApZ:EdKZ:CZ	5	9.5438	14	107.4549	0.0000	0.18305008	18.3050	79.4549	22.7586	0.0037	13	71.2264	76.3889
15*	IV:ApZ:EdJZ:CZ	5	9.5536	14	101.7185	0.0000	0.17327808	17.3278	73.7185	17.0222	0.0434	11	70.7547	76.3889
14	IV:ApZ:EdKZ:LZ	5	9.5540	14	101.4868	0.0000	0.17288344	17.2883	73.4868	16.7905	0.0737	13	71.6981	76.3889
13*	IV:ApZ:EdKZ	4	9.5628	12	96.2708	0.0000	0.16399792	16.3998	72.2708	23.6740	0.0274	10	70.5189	91.6667
12*	IV:ApZ:EdZ:CZ:KZ	4	9.5665	8	94.1440	0.0000	0.16037485	16.0375	78.1440	45.7461	0.0007	8	70.0472	76.3889
11*	IV:ApZ:EdZ:CZ:JZ	4	9.5756	8	88.7544	0.0000	0.15119366	15.1194	72.7544	40.3565	0.0301	9	69.5755	76.3889
10*	IV:ApZ:EdZ:KZ	3	9.5870	6	82.0689	0.0000	0.13980489	13.9805	70.0689	45.7705	0.0003	5	69.8113	91.6667
9*	IV:ApZ:EdZ:CZ	3	9.5908	5	79.8373	0.0000	0.13600335	13.6003	69.8373	49.5886	0.0010	6	69.5755	100.0000
8*	IV:ApZ:CZ:KZ	3	9.5910	6	79.7227	0.0000	0.13580807	13.5808	67.7227	43.4243	0.0009	5	68.6321	87.5000
7*	IV:ApZ:EdZ	2	9.6133	3	66.5901	0.0000	0.11343675	11.3437	60.5901	48.4409	0.0000	2	69.5755	100.0000
6*	IV:ApZ:CZ	2	9.6144	3	65.9506	0.0000	0.11234732	11.2347	59.9506	47.8014	0.0000	3	67.9245	100.0000
5*	IV:ApZ:KZ	2	9.6150	4	65.6244	0.0000	0.11179159	11.1792	57.6244	41.4255	0.0024	4	66.9811	100.0000
4*	IV:ApZ	1	9.6396	1	51.1652	0.0000	0.08716028	8.7160	49.1652	45.1155	0.0000	1	66.9811	100.0000
3*	IV:CZ	1	9.6998	2	15.7735	0.0004	0.02687026	2.6870	11.7735	3.6740	0.0004	1	58.0189	100.0000
2*	IV:EdZ	1	9.7009	2	15.1356	0.0005	0.02578359	2.5784	11.1356	3.0361	0.0005	1	56.3679	100.0000
1*	IV:Z	0	9.7266	0	0.0000	1.0000	0.00000000	0.0000	0.0000	0.0000	0.0000	0	52.1226	100.0000
ID	MODEL	Level	Н	dDF	dLR	Alpha	Inf	8dH (DV)	dAIC	dBIC	Inc.Alpha	Prog.	%C (Data)	<pre>%cover</pre>

Best	Model(s) by dBIC:													
9*	IV:ApZ:EdZ:CZ	3	9.5908	5	79.8373	0.0000	0.13600335	13.6003	69.8373	49.5886	0.0010	6	69.5755	100.0000
Best	Model(s) by dAIC:													
22	IV:ApZ:EdCZ:EdKZ:LZ	7	9.5221	20	120.2026	0.0000	0.20476589	20.4766	80.2026	-0.7921	0.0505	19	72.4057	52.7778
Best	Model(s) by Informat	ion, w	ith all	Inc.	Alpha < 0.	05:								
16*	IV:ApZ:EdKZ:CZ	5	9.5438	14	107.4549	0.0000	0.18305008	18.3050	79.4549	22.7586	0.0037	13	71.2264	76.3889 30

Search output (4/4)

Goodness measures for best models selected using BIC, AIC, p-value criteria:

- Reduction of DV uncertainty $\Delta H(DV | IV)$,
- Model complexity, df (degrees of freedom)
- Predictive accuracy: %correct

Search Output for dementia05.txt

Models selected by BIC/AIC/p-value (sort on info)

Criterion model %ΔH df %C BIC IV:ApZ:EdZ:CZ 5 14 70 14 71 p-value IV:ApZ:EdKZ:CZ 18 AIC IV:ApZ:EdCZ:EdKZ:LZ 20 20 72 SB BIC 3 $IV:Ap_0Ed_0Z:Ap_1Z:C_2Z:Z$ 14 70

IV = ApEdCKL... (all independent variables)

OCCAM Fit (VB, SB) input page

<u>F</u> il	e <u>E</u> dit <u>V</u> iev	v Hi <u>s</u> tory <u>B</u> o	okmarks <u>T</u> ools <u>H</u> elp Q S	earch				· _			x
J	Occam 3		× +					1			
ę	• • (•	🗲 🛈 dmit.	.sysc. pdx.edu /weboccam.cgi?ac	tion=fit&cached=	C 🟠	自合	1	→ ABP →	۵	• =	· 🔟
6	P	Portl	and State								
	OG	cam	version 3.3.11 — M	on Dec 19 21:06	:39 2016						
	O Do Se	arch	🔿 Do SB-Search	O Fit		🔘 Do SB-Fi	t	0 Do	Comp	are	
	© Show	Log	🗢 Manage Jobs					🗖 Cac	hed Da	ta Mode	
				Data Se	ettings						
	Data File:	:		Browse	No file select	ed.					
				Fit Set	ttings						-
	Model to	Fit:									
	Optional	default mode	el:								
	For direc	tod cystoms									
	Default	("negative")	DV state <i>(after rebinnin</i>	a) for confusion	matrices:	_					:
	Calcu	late expecte	ed DV								
	For pout	al evetome:									
	✓ Omit	table showir	ng all states for entire m	odel; 🗹 also omi	it tables for	IVI variable	S				
											-
				Hypergraph Di	splay Sett	ings	-14	O Curius			
	Hypergra	ph Layout St	tyle:	Fruchter Kamada-Kaw	man-keingo vai 🔘 Bipa	ntite	old- i liford	Sugiya	ama 🔍		
	🗹 Genera	ate graph im	ages								
	🗖 Genera	ate Gephi file	es								
	V Hide IV	V/IVI compor	nents								
	Use fu	ll variable na	ames in graph labels								_
				Output S	Settings						
	Return Gephi file	i data in CS\ s)	/ format (if hypergraph re	endering enabled	, return ZIF	o format con	taining CSV	, PDF, and	đ		
	🗹 Print o	ption setting	gs	🔲 (but don'i	t print varia	able definition	ns)				
	Run in Ba	ckground, Er	mail Results To:								
	Subject li	ine for email	(optional):							Send	

Fit output

- A model <u>is</u> a calculated conditional probability distribution, p_{model}(DV | IV),
 e.g., p_{ABC:AZ:BZ} (Z_I | A_i, B_j, C_k)
- Predict DV (Z) from IVs (A,B,C), using this conditional p distribution (stochastic)

Fit Output for dementia05.txt model IV:ApZ:EdZ:CZ (has loop) (demo#3)

Conditional DV (D) (%) for each IV composite state for the Model IV:ApZ:EdZ:CZ. IV order: ApEdC (APOE; Education; rs7561528).

IV			Т	Data			Т	Model						
			Т		obs. p	(DV IV)	Т	calc. o	(DV IV)					
Ap	Ed	С	1	freq	Z=0	Z=1	T.	Z=0	Z=1	rule	#correct	<pre>%correct</pre>	p(rule)	p(margin)
0	0	0	1	6.000	33.333	66.667	1	16.719	83.281	1	4.000	66.667	0.103	0.083
0	0	1	1	4.000	0.000	100.000	1	21.075	78.925	1	4.000	100.000	0.247	0.214
0	0	2	1	7.000	14.286	85.714	1	35.284	64.716	1	6.000	85.714	0.436	0.372
0	1	0	Т	8.000	50.000	50.000	Т	52.532	47.468	0	4.000	50.000	0.886	0.981
0	1	1	1	40.000	62.500	37.500	$^{+}$	59.546	40.454	0	25.000	62.500	0.227	0.347
0	1	2	1	44.000	72.727	27.273	Т	75.034	24.966	0	32.000	72.727	0.001	0.002
0	2	0	1	14.000	42.857	57.143	$^{+}$	55.306	44.694	0	6.000	42.857	0.691	0.811
0	2	1	- I	86.000	60.465	39.535	1	62.205	37.795	0	52.000	60.465	0.023	0.061
0	2	2	1	68.000	83.824	16.176	$^{+}$	77.067	22.933	0	57.000	83.824	0.000	0.000
1	0	0	Т	3.000	0.000	100.000	Т	3.984	96.016	1	3.000	100.000	0.111	0.095
1	0	1	1	5.000	20.000	80.000	Т	5.230	94.770	1	4.000	80.000	0.045	0.036
1	0	2	1	3.000	33.333	66.667	1	10.127	89.873	1	2.000	66.667	0.167	0.146
1	1	0	1	8.000	25.000	75.000	1	18.614	81.386	1	6.000	75.000	0.076	0.058
1	1	1	1	19.000	26.316	73.684	Т	23.325	76.675	1	14.000	73.684	0.020	0.012
1	1	2	1	21.000	33.333	66.667	Т	38.315	61.685	1	14.000	66.667	0.284	0.206
1	2	0	1	12.000	25.000	75.000	Т	20.366	79.634	1	9.000	75.000	0.040	0.028
1	2	1	1	40.000	25.000	75.000	Т	25.381	74.619	1	30.000	75.000	0.002	0.001
1	2	2	1	36.000	36.111	63.889	Т	40.986	59.014	1	23.000	63.889	0.280	0.181
			Т	424.000	52.123	47.877	Т	52.123	47.877	0	295.000	69.575		
			I.	freq	Z=0	Z=1	1	Z=0	Z=1	rule	#correct	<pre>%correct</pre>	p(rule)	p(margin)

* Rule selected using the independence model.

Input file for <u>SB</u>-search (for demo#4) dementia05ApEdC.txt: reduce #IVs to 3

#Fifth test set of OHSU data 20 June 2012
#Corrected transform of variable I rs12248379
#Corrected transform of APOE
#Updated imputed data
#Deleted var I and dropped records with no var C data
#Added stats on missing. Blank = none missing

:nominal					
ID	,0,0,ID				
APOE	,2,1,AP				
Gender	,2,0,5X				
Education	.3.1.ED				
AgeLastExam	.3.0.AG				
rš1801133	.3.0.A				
rs3818361	4.0.B	#	missing	3	
rs7561528	.3.1.C				
rs744373	,3,0,D				
rs6943822	3,0,E				
rs4298437	,3,0,F				
rs7012010	,3,0,G				
rs11136000	,3,0,H				
rs10786998	,4,0,3	#	missing	9	
rs11193130	,4,0,K	#	missing	11	
rs610932	,3,0,L		-		
rs3851179	,3,0,M				
rs3764650	,4,0,N	#	missing	2	
rs3865444	,4,0,P	#	missing	9	
CaseControl	,2,2,Z		-		

:no-frequency

:data #IndID	APOE	GENDER	EDU	ALE	А	в	с	D	E	F	G	н	J	к
101	0	0	2	2	1	1	0	1	2	2	1	1	2	0
103	0	0	2	1	0	2	2	0	1	1	1	2	2	0
111	0	1	2	1	2	2	1	1	0	1	1	2	1	1
112	0	0	2	2	2	2	1	1	1	2	1	1	0	2
118	0	1	0	2	2	2	2	0	0	1	1	1		
120	0	1	2	2	1	2	1	1	0	1	1	2	1	1
121	0	0	2	2	2	2	1	1	2	0	0	0	2	0
122	0	0	1	2	1	2	1	1	2	0	0	2	2	0
123	0	0	2	2	2	2	2	0	1	1	0	0	2	0
126	0	0	2	2	2	2	2	0	1	1	2	2	2	0
127	0	1	2	2	0	2	1	1	1	1	1	2	2	0
128	0	0	2	2	1	2	1	1	0	1	1	2		
129	1	1	2	2	0	2	1	0	1	0	1	2	0	2
132	0	0	2	2	2	2	1	0	0	1	1	2	1	1 20
134	0	1	2	2	0	0	2	0	0	1	1	1	0	2 30
135	0	0	2	2	2	1	1	0	0	2	1	1	0	2

SB search output (demo#4)

ID	MODEL	Level	H	dDF	dLR	Alpha	Inf	8dH (DV)	dAIC	dBIC	Inc.Alpha	Prog.	%C(Data)	<pre>%cover</pre>
16	IV:Ap0Ed0C0Z:Ap0Ed2C2Z:Ap0Ed0Z:Ap1Z:C2Z:Z	5	4.3897	5	86.7978	0.0000	0.95071864	14.7861	76.7978	56.5491	0.0949	12	69.5755	100.0000
15	IV:Ap0Ed0C0Z:Ap0Ed0Z:Ap0C0Z:Ap1Z:C2Z:Z	5	4.3906	5	86.2578	0.0000	0.94480434	14.6941	76.2578	56.0091	0.1337	12	70.0472	100.0000
14	IV:Ap0Ed2C2Z:Ap1Ed0C0Z:Ap0Ed0Z:Ap1Z:C2Z:Z	5	4.3906	5	86.2410	0.0000	0.94462025	14.6912	76.2410	55.9923	0.1962	13	69.5755	100.0000
13*	IV:Ap0Ed2C2Z:Ap0Ed0Z:Ap1Z:C2Z:Z	4	4.3935	4	84.5686	0.0000	0.92630198	14.4063	76.5686	60.3697	0.0337	9	69.5755	100.0000
12	IV:Ap0Ed0C0Z:Ap0Ed0Z:Ap1Z:C2Z:Z	4	4.3944	4	84.0051	0.0000	0.92012989	14.3103	76.0051	59.8062	0.1044	10	69.5755	100.0000
11	IV:Ap0Ed0Z:Ap0C2Z:Ap1Z:Ed2C2Z:Z	4	4.3951	4	83.6256	0.0000	0.91597252	14.2457	75.6256	59.4266	0.0567	8	69.5755	100.0000
10*	IV:Ap0Ed0Z:Ap1Z:C2Z:Z	3	4.3989	3	81.3650	0.0000	0.89121192	13.8606	75.3650	63.2158	0.0002	7	69.5755	100.0000
9*	IV:Ap0Ed2C2Z:Ap0Ed0Z:Ap1Z:Z	3	4.4011	3	80.0645	0.0000	0.87696704	13,6390	74.0645	61.9153	0.0002	6	69.5755	100.0000
8*	IV:Ap0Ed0Z:Ap0C2Z:Ap1Z:Z	3	4.4013	3	79.9909	0.0000	0.87616134	13.6265	73.9909	61.8417	0.0004	7	69.5755	100.0000
7*	IV:Ap0Ed0Z:Ap1Z:Z	2	4.4216	2	68.0475	0.0000	0.74534205	11.5919	64.0475	55.9480	0.0000	4	69.5755	100.0000
6*	IV:Ap0Ed2C2Z:Ap1Z:Z	2	4.4232	2	67.1062	0.0000	0.73503174	11.4316	63.1062	55.0067	0.0000	2	66.9811	100.0000
5*	IV:Ap1Z:Ed0Z:Z	2	4.4243	2	66.4783	0.0000	0.72815386	11.3246	62.4783	54.3788	0.0001	4	69.5755	100.0000
4*	IV:Ap1Z:Z	1	4.4504	1	51.0910	0.0000	0.55961325	8.7034	49.0910	45.0413	0.0000	1	66.9811	100.0000
3*	IV:Ap0C2Z:Z	1	4.4727	1	37.9865	0.0000	0.41607561	6.4710	35.9865	31.9367	0.0000	1	62.2642	100.0000
2*	IV:Ap0Ed2C2Z:Z	1	4.4770	1	35.4733	0.0000	0.38854880	6.0429	33.4733	29.4236	0.0000	1	58.7264	100.0000
1*	IV:Z	0	4.5374	0	0.0000	1.0000	0.00000000	0.0000	0.0000	0.0000	0.0000	0	52.1226	100.0000
ID	MODEL	Level	H	dDF	dLR	Alpha	Inf	8dH (DV)	dAIC	dBIC	Inc.Alpha	Prog.	<pre>%C (Data)</pre>	<pre>%cover</pre>
Best	Model(s) by dBIC:													

10*	IV:Ap0Ed0Z:Ap1Z:C2Z:Z	3	4.3989	3	81.3650	0.0000	0.89121192	13.8606	75.3650	63.2158	0.0002	7	69.5755	100.0000	
Best	Model(s) by dAIC:														
16	IV:Ap0Ed0C0Z:Ap0Ed2C2Z:Ap0Ed0Z:Ap1Z:C2Z:Z	5	4.3897	5	86.7978	0.0000	0.95071864	14.7861	76.7978	56.5491	0.0949	12	69.5755	100.0000	
Best	Model(s) by Information, with all Inc. Alg	pha <	0.05:												37
13*	IV:Ap0Ed2C2Z:Ap0Ed0Z:Ap1Z:C2Z:Z	4	4.3935	4	84.5686	0.0000	0.92630198	14.4063	76.5686	60.3697	0.0337	9	69.5755	100.0000	-

SB-search output

• Example of a best model from SB-Search:

Best model by BIC criterion IV:Z: Ap_0Ed_0Z : Ap_1Z : C_2Z (note interaction effect)

For this data, get about same uncertainty reduction & %correct, but 2 df simpler
In other data often get higher uncertainty reduction & %correct for about same df

Non-standard or enhanced RA

Non-standard

- Time series analysis
- Continuous DVs (not for this course)
- Set-theoretic data (not for this course)

Enhanced

- Validation (training-test data splits)
- Inter-method comparison

Time series analysis Can similarly do spatial (e.g., GIS) analysis



mask



original data (numbers label variables)

U	V	W	Х	Υ	Ζ
-	I	1	I	I	I
-	I	-	1	I	I
0	1	2	3	4	5
3	4	5	6	7	8
6	7	8	9	10	11

transformed data XYZ(t) = ABC(t)UVW(t) = ABC(t-1)

Enhanced RA (for directed systems)

(Totally optional for this course)

Validation: training/test data splits or 3-way splits; 5- or 10-fold validation

Comparison with other data mining methods e.g., Logistic Regression, Support Vector Machines, Bayesian Networks, ...

RA framework (1/2)

bold = typical RA use; **blue** = in OCCAM; **red** = other pgms

1. VARIABLE	nominal (discrete: binary or multi-valued)
	ordinal (discrete)
	quantitative (bin continuous values for IVs)
2. SYSTEM	directed (deterministic/stochastic) (supervised learning)
	neutral (unsupervised learning)
3. DATA	Information-theoretic (IRA)
	frequency/probability distribution
	function ('k-systems' & 'u-systems' RA)
	set-theoretic (SRA) mapping, relation

RA framework (2/2)

bold = typical RA use; **blue** = in OCCAM; **red** = other pgms

4. PROBLEM	reconstruction (decomposition)
	confirmatory
	exploratory
	exhaustive (look at all models)
	heuristic (search lattice of models)
	identification (composition)
5. METHOD	variable-based (VB)
	state-based (SB)
	latent variable-based (LVB)

Questions?

LOOKING FORWARD

TO AN EXCITING, CHALLENGING, &

PRODUCTIVE COURSE !!