A brief introduction to DIGRAM

The DIGRAM projects

Descriptive statistics

Contingency tables

Tests of conditional independence

Graphical models

Rasch models

Graphical log-linear Rasch models

Demo

Exercises

Creating DIGRAM projects

DIGRAM

is a program for discrete graphical modeling, including

tests of conditional independence in large sparse tables,

analysis of ordinal categorical date by high-dimensional chain graph models,

item analysis by graphical and log-linear Rasch models

and many other things that I have needed as an applied statistician

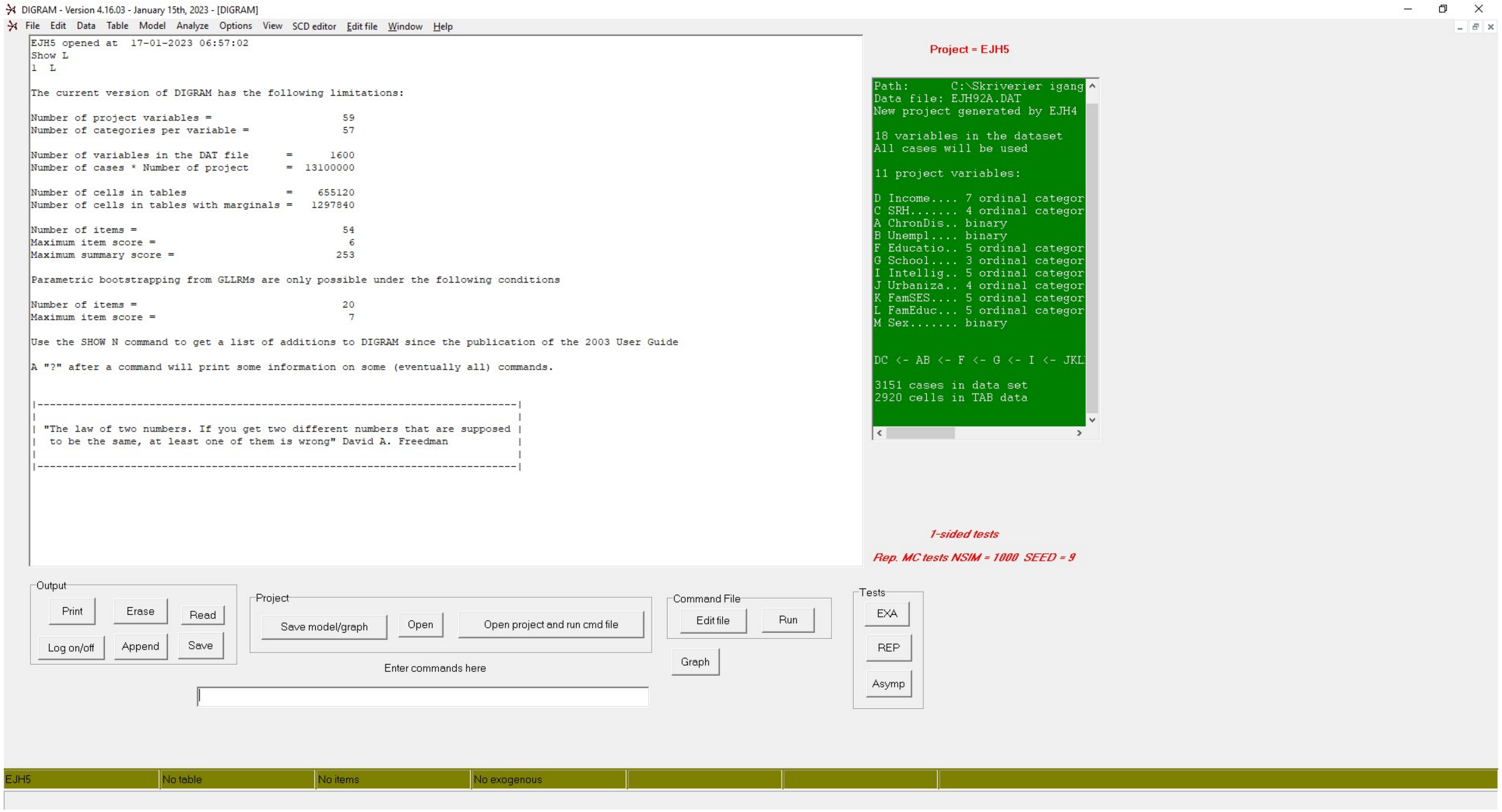
DIGRAM projects

Data for analysis by DIGRAM are organized in DIGRAM projects defined by the following files

File	Contents	Useful commands
Projectname.def	Info on where to find data	-
Projectname.var	Definition of variables	VAR to redefine variables
Projectname.cat	Definition of categories	CAT to redefine categories
Projectname.dat	The original data set	-
Projectname.sys	Data with categorized variables	MAKE to create sys and tab files from data
Projectname.tab	Tabulated data	
Projectname.grf	Definition of graphical project model	SHOW G to show the definition of the model

Invoking DIGRAM

Opens the previous DIGRAM project and summarize limitations of the current version of the program



The EJH5 project – data from a Danish panel study

A “SHOW V” command provide info on the project variables

```
+-------+
|       |
| EJH5  |
|       |
+-------+
D:    Income  -   7 ordinal categories
C:       SRH  -   4 ordinal categories
A:  ChronDis  -   2 ordinal categories
B:    Unempl  -   2 ordinal categories
F:  Educatio  -   5 ordinal categories
G:    School  -   3 ordinal categories
I:  Intellig  -   5 ordinal categories
J:  Urbaniza  -   4 ordinal categories
K:    FamSES  -   5 ordinal categories
L:   FamEduc  -   5 ordinal categories
M:       Sex  -   2 ordinal categories

CAUSAL/RECURSIVE STRUCTURE
D,C <- A,B <- F <- G <- I <- J,K,L,M
```

L	FamEduc	M	Sex
1	Long	1	Male
2	Med+Shor	2	Female
3	Vocation		
4	Training		
5	None		

D	Income	C	SRH	A	ChronDis
1	< 100	1	VeryGood	1	None
2	100-200	2	Fair	2	OneOrMor
3	150-200	3	LessFair		
4	200-250	4	Bad		
5	250-300				
6	300-400				
7	400+				

B	Unempl	F	Educatio	G	School
1	< 1 year	1	LOng	1	7 years
2	1+ years	2	Medium	2	8-10
		3	Short	3	HighScho
		4	Vocation		
		5	None		

I	Intellig	J	Urbaniza	K	FamSES
1	-25	1	Copenhag	1	I (high)
2	26-30	2	City+Tow	2	II
3	31-35	3	Village	3	III
4	36-40	4	Countrys	4	IV
5	41+			5	V (Low)

Descriptive statistics

Correlations in DIGRAM are measured by Goodman & Kruskall's Gamma for ordinal categorical variables

FREQ I

```
+-------------+
|             |
| I: Intellig |
|             |
+-------------+

Reference no.    7
Variable no.    17

            I  Count      Pct   CumPct
--------------------------------------
      -25   1    359    11.39    11.39
    26-30   2    410    13.01    24.40
    31-35   3    635    20.15    44.56
    36-40   4    720    22.85    67.41
      41+   5   1027    32.59   100.00

TOTAL   3151

===================================
```

COR DFGIKM

	Income	Educatio	School	Intellig	FamSES	Sex
Income		-0.3882	0.2836	0.1982	-0.1760	-0.6048
p		0.0000	0.0000	0.0000	0.0000	0.0000
Educatio	-0.3882		-0.6853	-0.4283	0.3542	0.0187
p	0.0000		0.0000	0.0000	0.0000	0.5277
School	0.2836	-0.6853		0.5652	-0.4777	0.1600
p	0.0000	0.0000		0.0000	0.0000	0.0000
Intellig	0.1982	-0.4283	0.5652		-0.2532	0.0285
p	0.0000	0.0000	0.0000		0.0000	0.2712
FamSES	-0.1760	0.3542	-0.4777	-0.2532		0.0313
p	0.0000	0.0000	0.0000	0.0000		0.2637
Sex	-0.6048	0.0187	0.1600	0.0285	0.0313	
p	0.0000	0.5277	0.0000	0.2712	0.2637	

Tabulating

Use TABULATE <variables> to generate contingency tables and SHOW T to see the table

```
 +-----------------------------+
 |                             |
 | Report on missing responses |
 |                             |
 +-----------------------------+

D   Income  Observed =   2342 Missing = 809
I Intellig  Observed =   3151 Missing =   0
M      Sex  Observed =   3151 Missing =   0

Number of cases with complete responses =
2342 out of 3151

Distribution of number of missing variables

     Count
----------
 0    2342
 1     809
 2       0
 3       0
----------
```

```
 +-------------------------------------+
 |                                     |
 | Variables included in the current table |
 |                                     |
 +-------------------------------------+

D:   Income  -   7 ordinal categories
I: Intellig  -   5 ordinal categories
M:      Sex  -   2 ordinal categories

CAUSAL/RECURSIVE STRUCTURE
D <- I <- M

                The DIM table.
```

M	I	D=1	D=2	D=3	D=4	D=5	D=6	D=7
1	1	7	14	41	38	18	8	5
	2	4	6	45	60	23	17	1
	3	8	14	56	73	28	25	12
	4	11	16	47	65	35	40	14
	5	9	13	48	95	77	79	62
2	1	15	28	41	18	2	1	0
	2	19	49	54	21	3	4	0
	3	29	86	91	40	12	3	2
	4	24	92	101	72	9	12	2
	5	34	79	112	118	32	17	6

Tests of conditional independence

HYP DI **defines the hypothesis of conditional independence of D (income) and I (intelligence) given M (sex)**

```
1 Hypothesis:

HYPOTHESIS  1:  D & I  |  M
```

Test T **shows the test statistics and the table**

Sex (M)	Intellig (I)	< 100	100-2	150-2	200-2	250-3	300-4	400+	TOTAL	
		D:--Income								
1	-25	7	14	41	38	18	8	5	131	
	row%	5.3	10.7	31.3	29.0	13.7	6.1	3.8	100.0	
	26-30	4	6	45	60	23	17	1	156	
	row%	2.6	3.8	28.8	38.5	14.7	10.9	0.6	100.0	
	31-35	8	14	56	73	28	25	12	216	
	row%	3.7	6.5	25.9	33.8	13.0	11.6	5.6	100.0	
	36-40	11	16	47	65	35	40	14	228	
	row%	4.8	7.0	20.6	28.5	15.4	17.5	6.1	100.0	
	41+	9	13	48	95	77	79	62	383	
	row%	2.3	3.4	12.5	24.8	20.1	20.6	16.2	100.0	X^2 = 120.7
										df = 24
	TOTAL	39	63	237	331	181	169	94	1114	p = 0.000
	row%	3.5	5.7	21.3	29.7	16.2	15.2	8.4	100.0	Gam = 0.29
										p = 0.000
2	-25	15	28	41	18	2	1	0	105	
	row%	14.3	26.7	39.0	17.1	1.9	1.0	0.0	100.0	
	26-30	19	49	54	21	3	4	0	150	
	row%	12.7	32.7	36.0	14.0	2.0	2.7	0.0	100.0	
	31-35	29	86	91	40	12	3	2	263	
	row%	11.0	32.7	34.6	15.2	4.6	1.1	0.8	100.0	
	36-40	24	92	101	72	9	12	2	312	
	row%	7.7	29.5	32.4	23.1	2.9	3.8	0.6	100.0	
	41+	34	79	112	118	32	17	6	398	
	row%	8.5	19.8	28.1	29.6	8.0	4.3	1.5	100.0	X^2 = 73.0
										df = 24
	TOTAL	121	334	399	269	58	37	10	1228	p = 0.000
	row%	9.9	27.2	32.5	21.9	4.7	3.0	0.8	100.0	Gam = 0.21
										p = 0.000

The Chi-squared test of conditional independence is the sum of the Chi-squared statistics in the different strata.

The partial gamma coefficient is a weighted average of the gamma coefficients in the different strata.

p-values are repeated Monte Carlo estimates of exact conditional p-values

```
****  Summary of results  ****

NSIM = 1000 tables generated for exact p-values

single_exa_summary under reconstruction
-------------------------------------------------------------------------------------
                         p-values                         p-values (1-sided)
Hypothesis        X²  df asymp exact 99% conf.int. Gamma asymp exact 99% conf.int. nsim
-------------------------------------------------------------------------------------
 1:D&I|M      193.7  48 0.000 0.000 0.000 - 0.007  0.24 0.000 0.000 0.000 - 0.007 1000 xx ++
-------------------------------------------------------------------------------------
-----------------------------------------------------------
** Local testresults for strata defined by      Sex  (M) **
                          p-values      p-values (1-sided)
 M:      Sex    X²    df asympt  exact  Gamma asympt  exact
-----------------------------------------------------------
 1:     Male 120.70   24 0.0000 0.0000   0.29 0.0000 0.0000
 2:   Female  73.00   24 0.0000 0.0000   0.21 0.0000 0.0000
-----------------------------------------------------------
```

Income at age 43 is conditionally dependent of Intelligence given Sex

Hyp DM and TEST L creates the hypothesis of conditional independence of D and M and calculates overall and local test statistics but do not show the table.

Hypothesis	X^2	df	p-values asymp	p-values exact	99% conf.int.	Gamma	p-values (1-sided) asymp	p-values (1-sided) exact	99% conf.int.	nsim	
1:D&M|I	533.7	30	0.000	0.000	0.000 - 0.007	-0.63	0.000	0.000	0.000 - 0.007	1000	**xx** --

** Local testresults for strata defined by Intellig (I) **

I:	Intellig	X^2	df	p-values asympt	p-values exact	Gamma	p-values (1-sided) asympt	p-values (1-sided) exact
1:	-25	35.53	6	0.0000	0.0000	-0.54	0.0000	0.0000
2:	26-30	87.34	6	0.0000	0.0000	-0.71	0.0000	0.0000
3:	31-35	109.00	6	0.0000	0.0000	-0.63	0.0000	0.0000
4:	36-40	107.34	6	0.0000	0.0000	-0.57	0.0000	0.0000
5:	41+	194.49	6	0.0000	0.0000	-0.65	0.0000	0.0000

Income at age 43 is conditionally dependent of sex given Intelligence

Will it change anything if we also stratify with education

Income (D) and Intelligence (I) given Education (F) and Sex (M)

TAB DFIM creates the table

Hyp DI defines the hypothesis

LOCAL asks for local test results for

Test tests the hypothesis but does not print the table

Significant evidence of weak effect of intelligence at age 18 on income at age 43

Hypothesis	X^2	df	p-values asymp	p-values exact	99% conf.int.	Gamma	p-values (1-sided) asymp	p-values (1-sided) exact	99% conf.int.	nsim	
1:D&I|FM	254.0	236	0.201	0.251	0.217 - 0.288	0.10	0.000	0.000	0.000 - 0.007	1000	++

Local test results in subpopulations defined by education or sex

** Local testresults for strata defined by Educatio (F) **

F:	Educatio	X^2	df	p-values asympt	p-values exact	Gamma	p-values (1-sided) asympt	p-values (1-sided) exact
1:	LOng	53.95	48	0.2574	0.2910	0.18	0.0677	0.0930
2:	Medium	36.81	48	0.8802	0.8280	-0.03	0.3217	0.3360
3:	Short	65.34	48	0.0485	0.0510	0.03	0.3068	0.3130
4:	Vocation	51.97	48	0.3221	0.3240	0.12	0.0001	0.0000
5:	None	45.92	44	0.3924	0.4040	0.07	0.0768	0.0700

** Local testresults for strata defined by Sex (M) **

M:	Sex	X^2	df	p-values asympt	p-values exact	Gamma	p-values (1-sided) asympt	p-values (1-sided) exact
1:	Male	120.75	120	0.4636	0.4470	0.16	0.0000	0.0000
2:	Female	133.24	116	0.1306	0.1610	0.04	0.1099	0.1080

The effect of intelligence on income appears to be a local phenomenon.

Evidence of effect is only found for males or persons with either vocational education

The methodological issue

Would our results change if we used or included different control variables?

And if they do, how do we select an appropriate set of control variables?

Graphical models used in the right way may provide the answers to these questions.

Graphical models

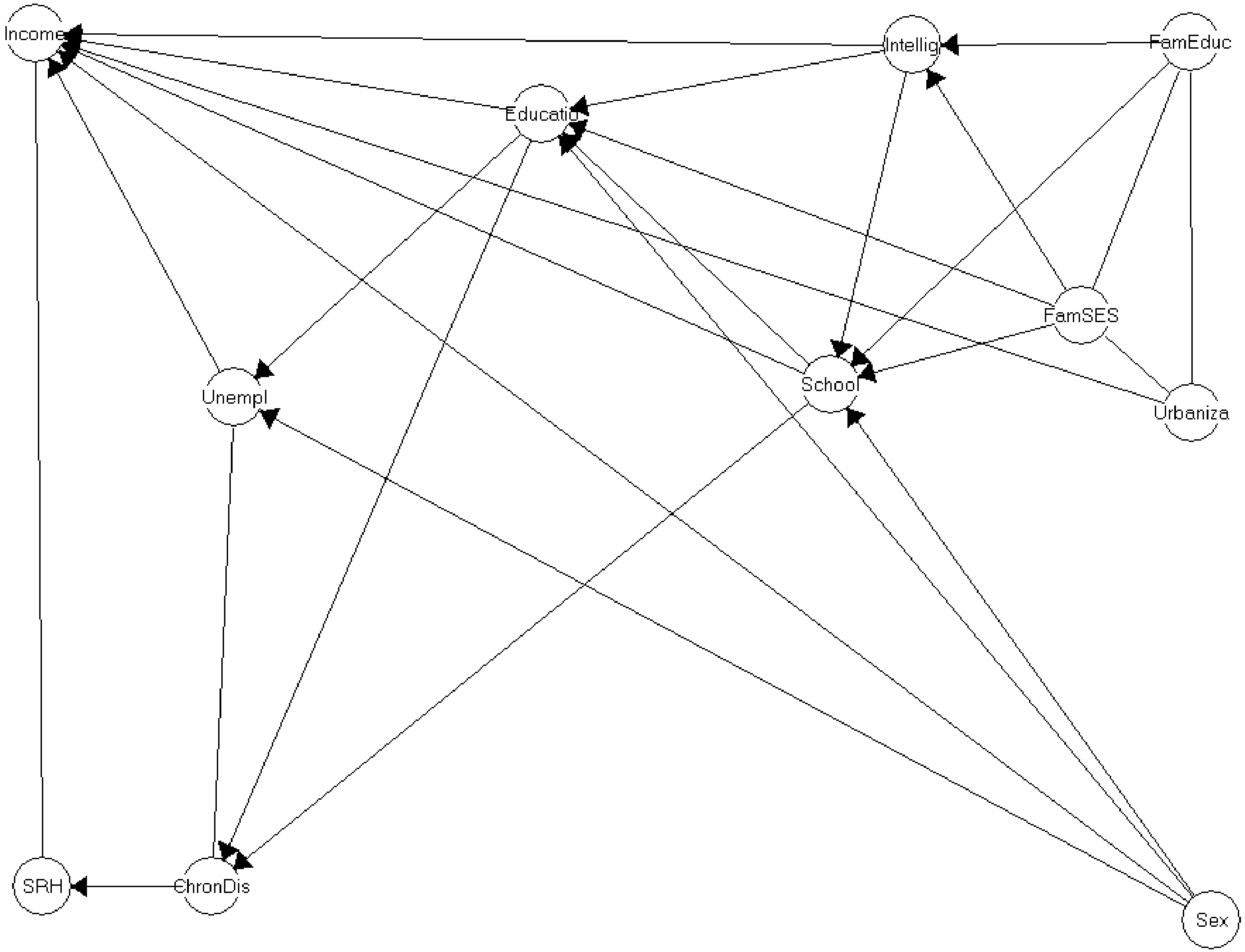
Block recursive graphical models are defined by recursive structure and assumptions of conditional independence of variables given all concurrent and prior variables.

The model is defined by Markov graphs with nodes representing variables and with missing edges and arrows between conditionally independent variables.

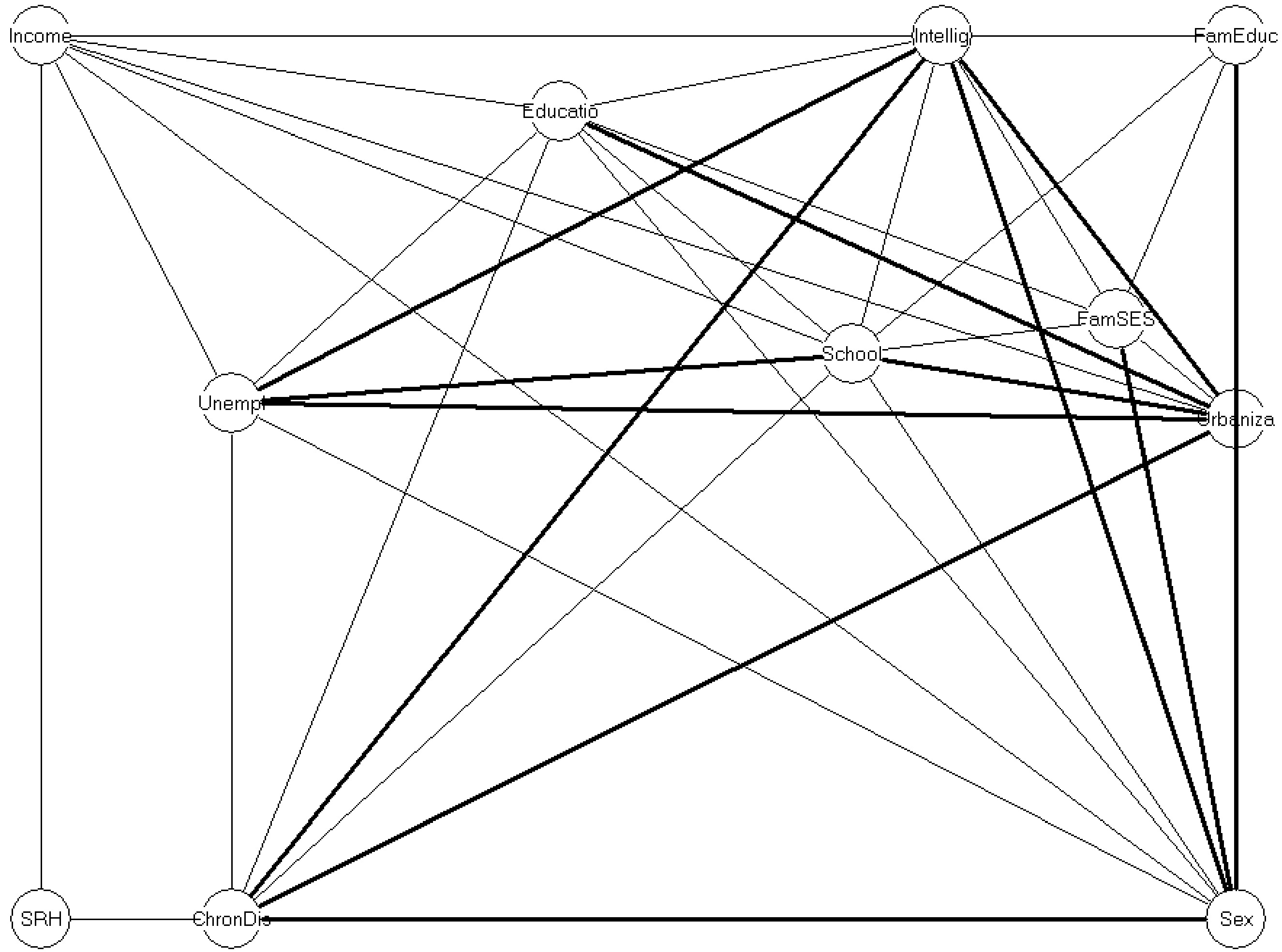
Moralization of graphical models define undirected graphical models that we can use to identify global Markov properties implying conditional independence given subsets of variables.

SEP and GMP commands define the variables needed for tests of conditional independence in graphical models.

The EJH5 model



The moral EJH5 graph – solid edges define associations that are needed for the graph to be moral



To test the dependence of income on intelligence we use the SEP command to define the hypothesis of conditional independence.

```
Separation hypotheses:

2 Hypotheses:

HYPOTHESIS  1:  D & I  |  C B F G J M
HYPOTHESIS  2:  D & I  |  A B F G J M
```

To test these hypotheses and the test of conditional independence given all variables we use the GTEST DI command

```
-----------------------------------------------------------------------------------------------
                        p-values    p-values (1-sided)     95% confidence
Hypothesis       X²  df asymp exact  Gamma   asymp exact       interval     nsim      n
-----------------------------------------------------------------------------------------------
 1:D&I|CBFGJM1329.51168 0.001 0.789   0.10   0.003 0.003     [0.03 - 0.17]  1000    1811     +
 2:D&I|ABFGJM1334.21195 0.003 0.805   0.08   0.012 0.020     [0.01 - 0.15]  1000    1895     +
 3:D&I|CABFGJ
   ---|M      1234.01078 0.001 0.940   0.09   0.009 0.014     [0.02 - 0.17]  1000    1644     +
-----------------------------------------------------------------------------------------------
```

The results confirm our finding. Intelligence has a weak effect on income.

Tests of conditional independence during item analysis

Let # be the raw score over the ten PF items of SF-36. One way to test the hypothesis of no DIF of the third item (C) relative to sex (K) would be to test conditional independence of C and K given #.

The test provide very strong evidence of DIF.

Hypothesis	X^2	df	p-values asymp	p-values exact	99% conf.int.	Gamma	p-values (1-sided) asymp	p-values (1-sided) exact	99% conf.int.	nsim	
1:C&K|#	44.4	20	0.001	0.002	0.000 - 0.010	-0.52	0.002	0.001	0.000 - 0.008	1000	xx --

** Local testresults for strata defined by RawScore (#) **

#:	Score	X^2	df	p-values asympt	p-values exact	Gamma	p-values (1-sided) asympt	p-values (1-sided) exact
4:	3	0.75	1	0.3865	1.0000	-1.00	0.1103	0.6720
5:	4	6.00	1	0.0143	0.1010	-1.00	0.0000	0.0580
7:	6	1.74	1	0.1869	0.4510	-0.78	0.0828	0.0180
9:	7	0.75	1	0.3865	1.0000	-1.00	0.1103	0.6860
10:	9	2.06	1	0.1515	0.4240	-1.00	0.0352	0.2780
11:	10	2.25	1	0.1336	0.3410	1.00	0.1197	0.3410
12:	11	1.67	2	0.4346	0.5860	0.50	0.1702	0.4470
13:	12	0.95	2	0.6219	1.0000	0.33	0.2831	0.1510
14:	13	3.73	2	0.1551	0.2250	-0.80	0.0131	0.0080
15:	14	1.75	2	0.4164	0.4510	-0.46	0.1021	0.2060
16:	15	8.39	1	0.0038	0.0060	-0.91	0.0003	0.0000
17:	16	8.26	2	0.0161	0.0140	-0.88	0.0002	0.0050
18:	17	5.96	1	0.0147	0.0250	-0.82	0.0011	0.0180
19:	18	0.09	1	0.7632	1.0000	-0.14	0.3805	0.2320
20:	19	0.01	1	0.9259	1.0000	-0.07	0.4631	0.7370

Test of local independence by tests of conditional independence

To test that items D (walking two flights of stairs) is locally independent of E (one flight of stairs) we may test that the two items are conditionally independent given the rest-score (#) without D.

The test rejects local independence. The partial gamma coefficient shows that the local dependence is extremely strong.

Hypothesis	X^2	df	p-values asymp	p-values exact	99% conf.int.	Gamma	p-values (1-sided) asymp	p-values (1-sided) exact	99% conf.int.	nsim	
1:D&E|#	35.1	18	0.009	0.006	0.002 - 0.016	0.75	0.000	0.000	0.000 - 0.007	1000	**xx ++**

** Local testresults for strata defined by RawScore (#) **

#:	Score	X^2	df	p-values asympt	p-values exact	Gamma	p-values (1-sided) asympt	p-values (1-sided) exact
9:	8	0.24	1	0.6242	1.0000	1.00	0.1932	0.8390
10:	9	3.68	4	0.4510	0.6080	0.55	0.1459	0.1620
11:	10	4.55	2	0.1028	0.1340	-0.11	0.4383	0.4760
12:	11	3.81	2	0.1489	0.2120	1.00	0.0043	0.0900
13:	12	4.65	2	0.0978	0.1070	0.67	0.0082	0.0220
14:	13	4.79	2	0.0912	0.0980	0.68	0.0364	0.0290
15:	14	1.95	1	0.1625	0.1980	0.64	0.0603	0.1640
16:	15	2.53	1	0.1114	0.2200	1.00	0.0696	0.2200
17:	16	1.94	1	0.1634	0.3620	1.00	0.1541	0.3620

Invoke the IMP command to create a DIGRAM project

You need a csv file with data

During import, you have to decide on a recursive structure and how to categorize variables

Time for a little DEMO

SPSS data on happiness in Denmark in 2008

	N	Minimum	Maximum
SocSec *confidence, social security system	980	1	4
HealthCare *confidence, health care system	1003	1	4
Happy * happiness	1017	1	4
LifeSat * life satisfaction	1015	1	10
JobSat *job satisfaction	640	1	10
Control *control over life	1007	1	10
CivService *confidence, civil service	978	1	4
DemSat *satisfaction with democracy	949	1	4
Partner *stable relationship	1023	1	2
Sex *sex respondent	1023	1	2
Religion *Religious services	1017	1	8
Job *job	1023	1,00	5,00
Income *income household	908	1	10
Age	1023	17,00	91,00

A recursive structure for an analysis of happiness

(Happiness, LifeSat, Control over Life) ⇐ (income, Partner) ⇐ (Age, Sex)

Happiness

	n	%
1 very happy	459	44,9
2 quite happy	504	49,3
3 not very happy	50	4,9
4 not at all happy	4	,4
Total	1017	99,4

Satisfaction and control over life will be categorized

1-5 6-7 8 9 10

Satisfaction with life

	n	%
1 dissatisfied	9	,9
2	5	,5
3	12	1,2
4	19	1,9
5	55	5,4
6	47	4,6
7	89	8,7
8	249	24,3
9	239	23,4
10 satisfied	291	28,4
Total	1015	99,2

Control over life

	n	%
1 not at all	10	1,0
2	10	1,0
3	23	2,2
4	24	2,3
5	123	12,0
6	93	9,1
7	151	14,8
8	318	31,1
9	131	12,8
10 a great deal	124	12,1
Total	1007	98,4

Income

	n	%
1	90	8,8
2	92	9,0
3	87	8,5
4	109	10,7
5	74	7,2
6	94	9,2
7	83	8,1
8	93	9,1
9	88	8,6
10	98	9,6
Total	908	88,8

Partner

	n	%
1 yes	722	70,6
2 no	301	29,4
Total	1023	100,0

Sex

	n	%
1 male	504	49,3
2 female	519	50,7
Total	1023	100,0

Age

17 -91

Income categories

1-2 3-4 5-6 7-8 9-10

Age categories

17-29 30-39 40-49 50-59 60-60 70+