10-1998

Complexity Reduction in State-based Modeling

Martin Zwick
Portland State University, zwick@pdx.edu

COMPLEXITY REDUCTION IN STATE-BASED MODELING

Martin Zwick

Systems Science Ph.D. Program, Portland State University, OR

zwick@sysc.pdx.edu
http://www.sysc.pdx.edu/Faculty/Zwick

ABSTRACT

VARIABLE-BASED RECONSTRUCTABILITY ANALYSIS

Variables & Relations

Applications in Physical Systems

Specific & General Structures

Reconstructability Analysis

STATE-BASED RECONSTRUCTABILITY ANALYSIS (JONES)

The Basic Idea

Generalization

Examples

Open Questions

prepared for:
Session on *Dynamics and Complexity of Physical Systems*
International Conference on Complex Systems, Oct. 25-30, 1998

ABSTRACT

For a system described by a relation among qualitative variables (or quantitative variables "binned" into symbolic states), expressed either set-theoretically or as a multivariate joint probability distribution, complexity reduction (compression of representation) is normally achieved by modeling the system with projections of the overall relation. To illustrate, if ABCD is a four variable relation, then models ABC:BCD or AB:BC:CD:DA, specified by two triadic or four dyadic relations, respectively, represent simplifications of the ABCD relation. Simplifications which are lossless are always preferred over the original full relation, while simplifications which lose constraint are still preferred if the reduction of complexity more than compensates for the loss of accuracy.

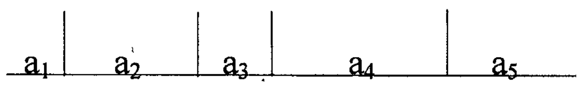
State-based modeling is an approach introduced by Bush Jones, which significantly enhances the compression power of information-theoretic (probabilistic) models, at the price of significantly expanding the set of models which might be considered. Relation ABCD is modeled not in terms of the projected relations which exist between subsets of the variables but rather in terms of a set of specific *states* of subsets of the variables, e.g., (A_i, B_j, C_k), (C_l, D_m), and (B_n). One might regard such state-based, as opposed to variable-based, models as utilizing an "event"- or "fact"-oriented representation. In the complex systems community, even variable-based decomposition methods are not widely utilized, but these state-based methods are still less widely known. This talk will compare state- and variable-based modeling, and will discuss open questions and research areas posed by this approach.

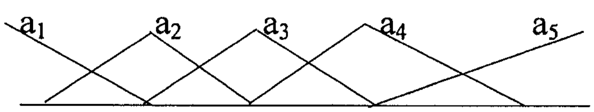
VARIABLE-BASED RECONSTRUCTABILITY ANALYSIS (RA)

Variables & Relations

1. **Nominal** state variables, e.g., $A = \{ a_1, a_2, a_3, \dots a_n \}$

Quant. var. with *non-linear* relations **binned:** *crisp* or *fuzzy* bins





2. State var. **sampled** by *support* variables (space, time, popul.)

E.g., in time-series analysis:

sv	...	t-2	t-1	t
U		G	D	A
V		H	E	B
W		I	F	C

3. **Relations** ($ABC \equiv R_{abc}$) are

(a) *directed* $\overset{A}{\underset{B}{\Rightarrow}} \boxed{ABC} \xrightarrow{C}$ (b) *neutral* $\overset{A}{\underset{B}{=}} \boxed{ABC} \overset{C}{\text{—}}$

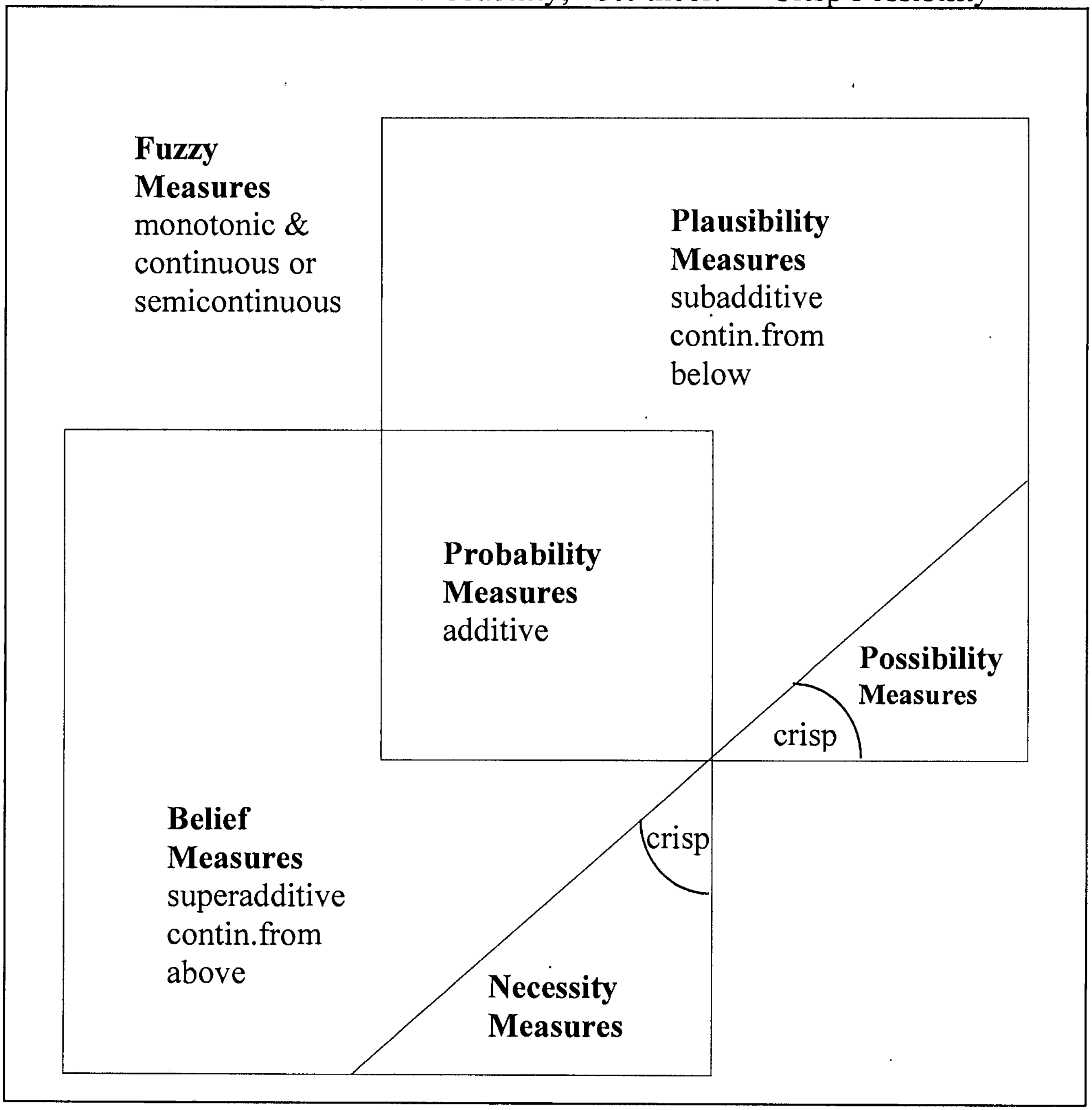
(a) *set*-theoretic (b) *info*.-theoretic (c) other

$ABC \subseteq A \otimes B \otimes C$
$= \{a_i, b_j, c_k\}$ *not all* ijk

$ABC = \{ p(a_i, b_j, c_k) \}$

(Klir)

“Information-theor.” = Probability; “Set-theor.” = Crisp Possibility



From **George J. Klir & Mark J. Wierman**, *Uncertainty-Based Information: Elements of Generalized Information Theory.* Springer-Verlag, 1998, p.40 (Figure 2.3. Inclusion relationships among relevant types of fuzzy measures.)

Potential Applications in Physical Systems

For *nominal* variables or if simulation of *non-linear quantitative* relations is difficult

1. Time series analysis; dynamic systems

- Chaotic vs. stochastic dynamics can be distinguished by info.-theor. analysis (Fraser)
- Chaos in cellular automata is predicted by RA (Zwick)
- Potential extension of RA analysis to continuous systems.
- (MacAuslan:) Nominal treatment of attractors, perhaps in weather modeling?

2. Other uses of nominal variables

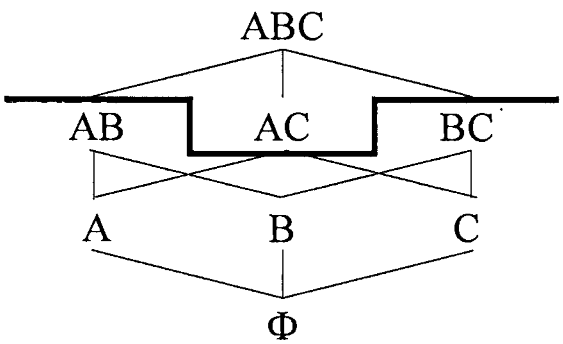
- Where quant. specification too detailed, e.g., amino acid types
- (MacAuslan:) Quantum states?

3. Where state-based methods might particularly apply

- Where features intrinsically multi-variate, perhaps image compression?
- Problems in high-dimension problems and sparse data

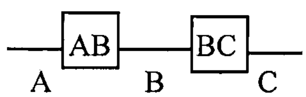
Specific and General Structures

1. Lattice of **Relations** (projections)

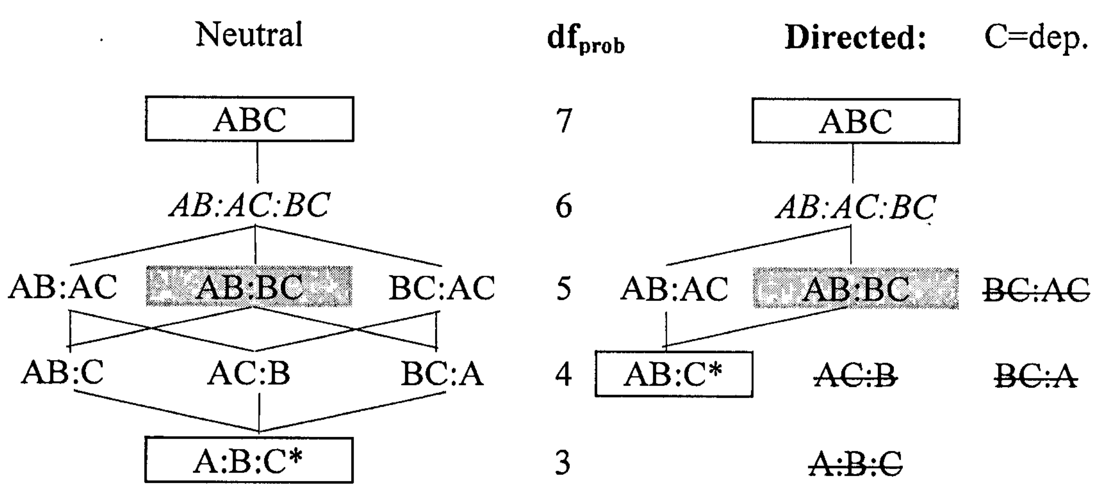


$\Phi_{\text{information-theoretic}}$ = uniform distribution

2. **Structure** = **cut** (above) through Lat. of Relations, e.g., AB:BC



3. Lattice of **Specific Structures** (*italics* = loops; ▭ = reference)



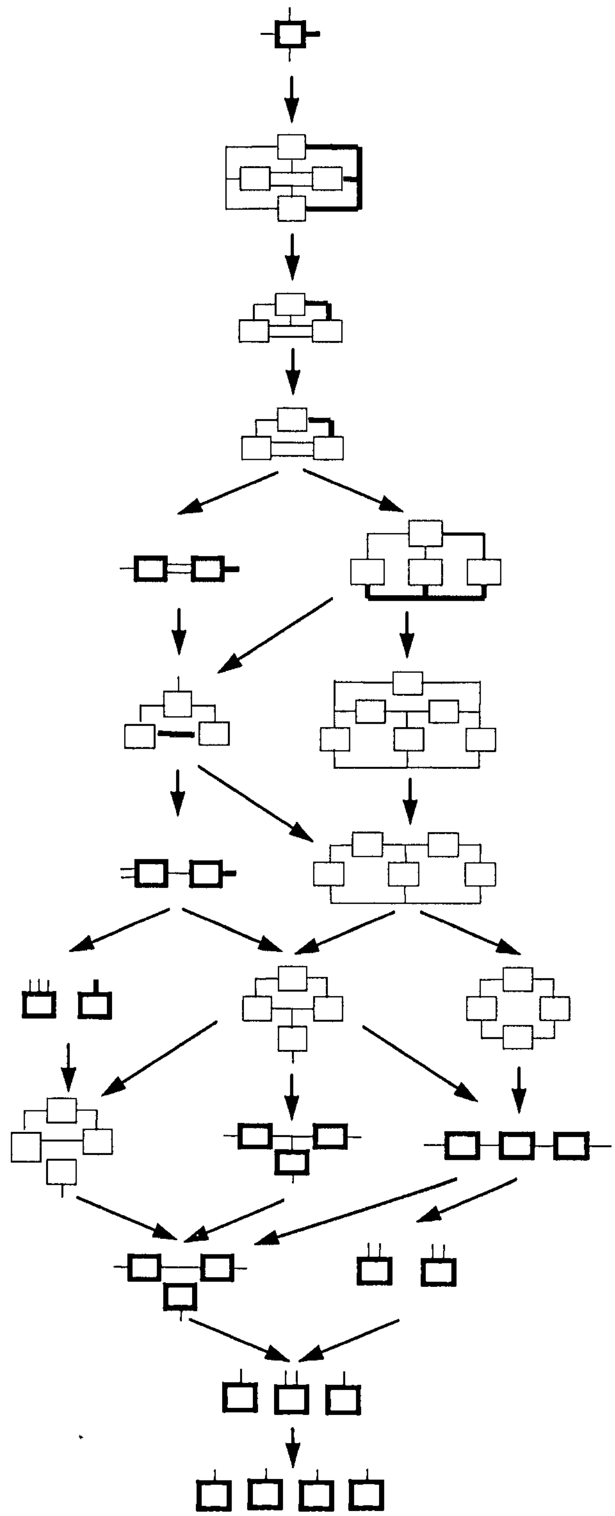
*Could extend Lattice down to Φ

Complexity = **df** = degrees of freedom given for binary variables

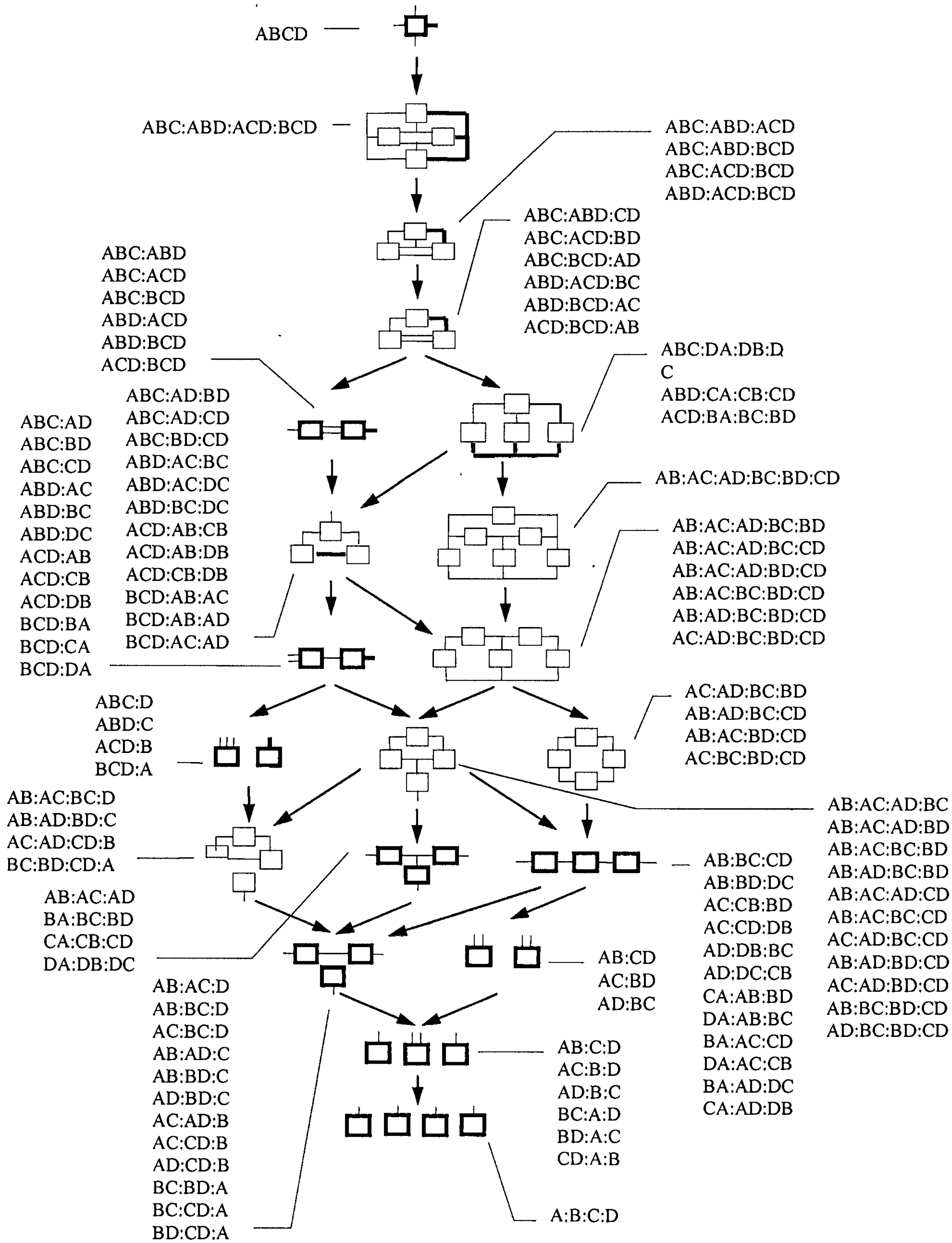
% complexity(AB:BC) = .5

4. Lattice of (20) **General Structures** for 4 variables.

Acyclic, directed structures indicated (1 dep. var.).



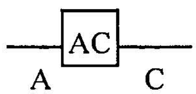
5. Four-Variable **Structures** (20 General, 114 Specific)



Complexity reduction with latent variables

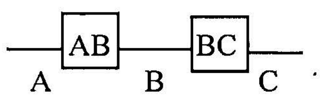
(*Factor analysis* for *nominal* variables)

Simplifying AC, with df(A)=df(C)=4 & **df(AC) = 15**,

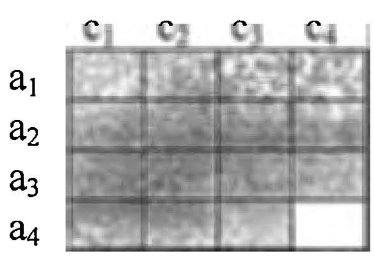


by **adding variable**, B, with df(B) = 2, & solving for an ABC

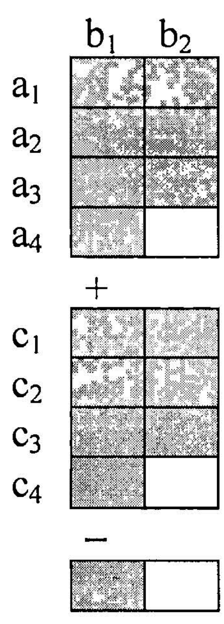
decomposable into **AB:BC,**



with **df(AB:BC)** = df(AB) + df(BC) - df(B) = 7 + 7 - 1 = **13**



AC



AB:BC

Reconstructability Analysis

1. **Constraint** *lost* and *retained* in structures.

ABC

T(AB:BC) = const. *lost* in AB:BC

AB:BC

T(A:B:C)

T(A:B:C)-T(AB:BC) = const. *captured* in AB:BC

A:B:C (or some other **reference structure**)

$T(AB:BC) = -\sum\sum\sum p(A,B,C) \log [\, p(A,B,C)/q_{AB:BC}(A,B,C) \,].$

$I = \%\ \text{info}_{\text{retained}} = [\ T(A:B:C) - T(AB:BC)\ ]\ /\ T(A:B:C)$

2. Models **lossless vs. lossy** in constraint

lossless: T = 0 (exactly or statistically); **lossy**: satisfice on I

statistical considerations: cut-offs for **Types I & II errors**

Top-down or bottom-up search:

descend lattice if constraint *lost* (T) is stat. insignificant or **small**

ascend lattice if constraint *retained* is stat. significant or **large**

3. Calculation of **model probabilities** (q's)

used in $T(A{:}B) = -\sum\sum p(A,B) \log [\, p(A,B) / \mathbf{q_{A:B}(A,B)} \,]$.

<u>Simpler example:</u>

	b_1	b_2	
a_1	.1	.2	.3
a_2	.3	.4	.7
	.4	.6	

observed $p(A,B)$; df 3

	b_1	b_2	
	q_{11}	q_{12}	.3
	q_{21}	q_{22}	.7
	.4	.6	

calculated $q_{A:B}(A,B)$; df 2

$q_{A:B}(A,B)$ is solution to:

maximize unc. $= -q_{11}\log q_{11} - q_{12}\log q_{12} - q_{21}\log q_{21} - q_{22}\log q_{22}$

subject to **linear constraints** of model, A:B :

complete margins: (model parameters)

$q_{11} + q_{12} = .3$

$\cancel{q_{21} + q_{22} = .7}$ (redundant*)

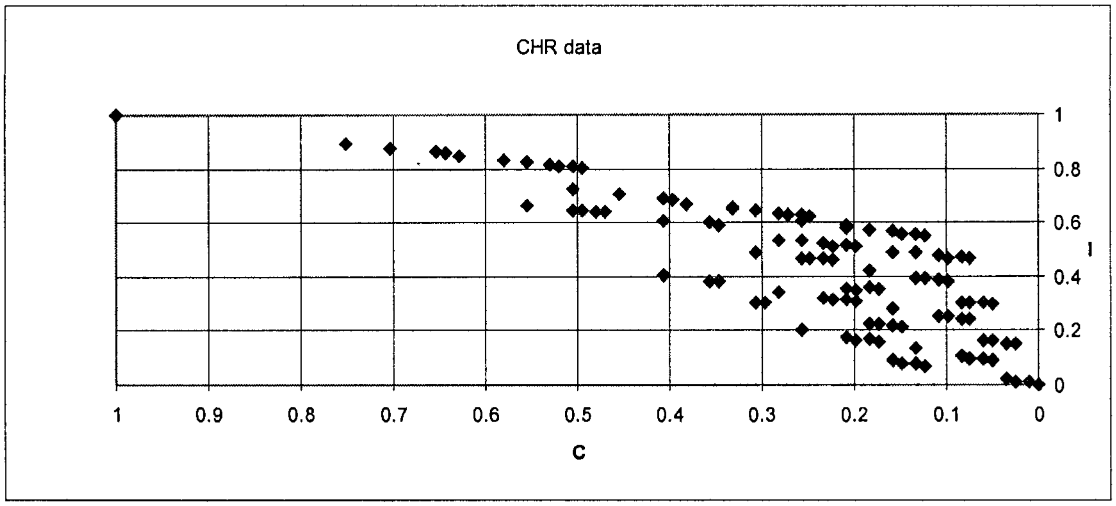
$q_{11} + q_{21} = .4$

$\cancel{q_{12} + q_{22} = .6}$ (redundant*)

*normalization: $q_{11} + q_{12} + q_{21} + q_{22} = 1$

Implemented by Iterative Proportional Fitting (IPF) algorithm

4. Example of examination of all 114 specific 4-var. structures



I = % information; C = % complexity

5. More variables ⇒ **combinatorial explosion**.

	number of structures			
# variables	3	4	5	6
# general structures	5	20	180	16,143
# specific structures	9	114	6,894	7,785,062
with 1 dep variable	5	19	167	7,580

Exhaustive search becomes impossible; need **heuristics**

1. **prune** tree as you go

2. **hierarchical** searching: coarse and fine searches

STATE-BASED RECONSTRUCTABILITY ANALYSIS (Bush Jones)

More powerful complexity reduction

The Basic Idea of SBRA

1. Simple example

.1	.1	.2
.1	.7	.8
.2	.8	

.04	.16	.2
.16	.64	.8
.2	.8	

q_{ij}

.1	*.1*
.1	**.7**

model	AB	A:B	a_2,b_2
df	3	2	1
loss	-	**.087**	**0**

$q_{a2,b2}(A,B)$ is solution to:

maximize unc. = $- q_{11} \log q_{11} - q_{12} \log q_{12} - q_{21} \log q_{21} - q_{22} \log q_{22}$

subject to **linear constraints**

of model, a_2,b_2: $q_{22} = .7$

& normalization: $q_{11} + q_{12} + q_{21} + q_{22} = 1$

(a_2, b_2) MODEL *SIMPLER* <u>**AND**</u> *MORE ACCURATE* THAN A:B

(Indeed, fits data perfectly!)

2. An interesting *supplementary* idea (Bush Jones):

(but for Jones, *inseparable* from SBRA.)

k-systems renormalization

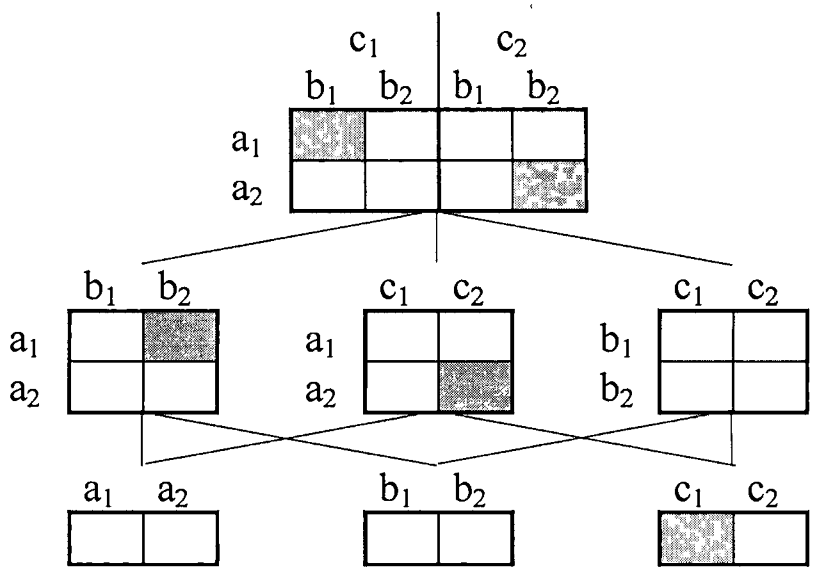
for SBRA of **arbitrary functions** of nominal variables

$$
\begin{tikzcd}
A \otimes B \to f(A,B) \arrow[d, "\begin{array}{l}\text{Renormalize f}\\ \text{to } [0,1] \text{ range}\\ \text{with } \sum = 1\end{array}"] & & \text{SBR of } f(A,B) \\
A \otimes B \to p(A,B) \arrow[rr, "\text{SBRA}"] & & \text{SBR of } p(A,B) \arrow[u, "\begin{array}{l}\text{Inverse}\\ \text{Normalization}\end{array}"']
\end{tikzcd}
$$

Generalization (**LOR** = lat. of relations; **LOS** = lat. of structures)

1. **Select** linearly-independent **set of states** from **LOR**

(**Variable**-based RA is a **special case** of **state**-based RA.)



2. LOS is **very big**! ⇒ **Stepwise state selection** heuristic (Jones):

1. $\mathbf{q^i}$, i=0, of **reference** = unif. distrib., Φ (bottom-up modeling)

2. ∀ candidate states, s, calculate *constraint captured* by state

$$I_s = p_s \log (p_s / q^i) + (1 - p_s) \log [(1 - p_s) / (1 - q^i)]$$

3. select **state** with **max. I**

4. **i ⇒ i+1, update q** by IPF for all states selected so far

5. go to 2

An Ecological Example

Analysis of algal productivity (Gary P. Shaffer)

	Factors	Value	Productivity	Information
1	Light Respiration Chlorophyl Tidal range	low low low high	14.6	52.48%
2	Light Chlorophyl Tidal range	high high low	53.0	93.27%
3	Light Chlorophyl Tidal range	high low low	30.2	98.39%
4	Light Respiration Chlorophyl Tidal range	high low low high	32.3	99.64%

Open Questions

1. Relation to **latent-variable** methods (replacing AC by A**B**:**B**C)

2. **Statistical significance** of added states, overall model

3. Relation to **ANOVA, non-hierarchical** log-linear methods

4. Improved LOS **search** algorithms (not sequential step)

5. **different reference** structures (not only Φ), e.g., A:B:C, AB:C

6. Use for **refining variable-based** RA

7. Extension to **set-theoretic** relations

8. Issues of **interpretation**

9. Validity of **k-systems** renormalization