SYSC 551 Discrete Multivariate Modeling
Prof. Martin Zwick
Portland State University
Notes from 2/02/2012
As taken by Juliana Arrighi & checked by MZ

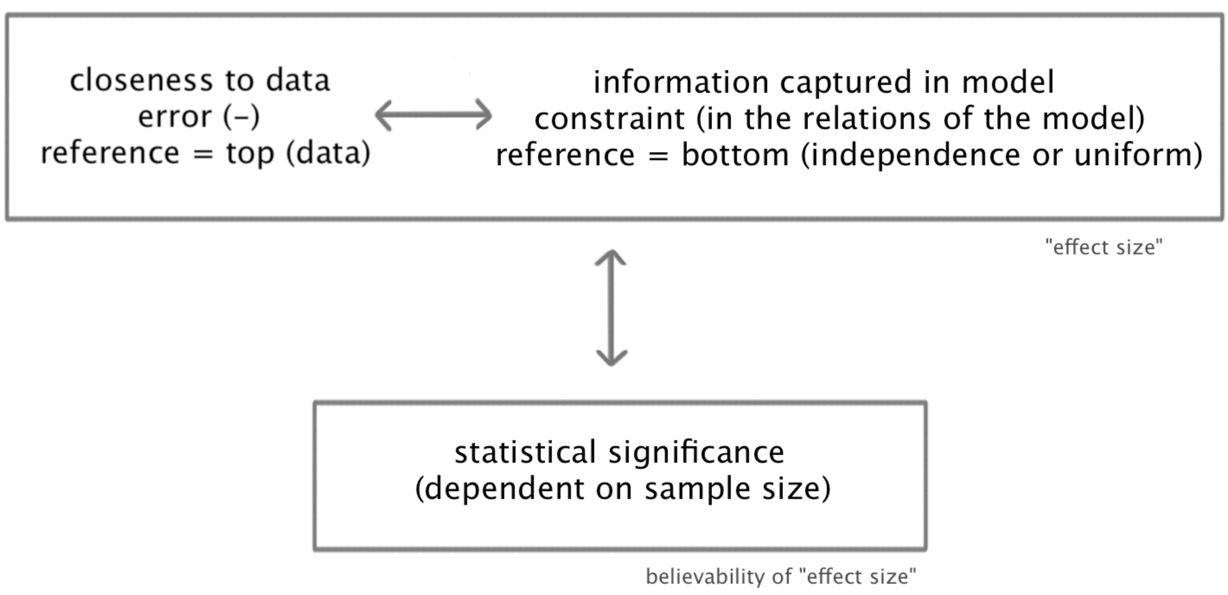
Information-Theoretic Reconstructability Analysis

(Putting it all together)

1. Preface: Goodness of a Model

We have talked about the goodness of a model being a trade-off between error (or its 'opposite,' information captured or effect size) and simplicity. We should also distinguish between effect size and statistical significance.

We then also have to consider statistical significance, which is the believability of these quantities, in other words, how certain we are that the effect size is not just due to chance. It is useful to examine both the effect size and its statistical significance. It is possible to have a small effect size that is highly significant, but that would not likely be much use to us.



2. Transmission and Information Distance

Reminder: the following are equivalent:

$T(x,y,z) = H(x) + H(y) + H(z) - H(x,y,z)$ variable notation

$T(X:Y:Z) = H(X) + H(Y) + H(Z) - H(XYZ)$ model notation

$T(m_j) = H(m_j) - H(m_o)$ Krippendorff notation where m_j = model, m_0 = data

The information distance between two models is the difference of their transmissions.

$$
\begin{aligned}
I(m_j \to m_k) &= T(m_k) - T(m_j) \\
&= H(m_k) - H(m_o) - \left(H(m_j) - H(m_o)\right) \\
&= H(m_k) - H(m_j)
\end{aligned}
$$

$$T(m_j) = \sum p(m_o)\log\left(\frac{p(m_o)}{q(m_j)}\right)$$

e.g. q(AB:BC) is calculated distribution for AB:BC
= p(AB)p(C|B)
= p(AB)p(BC)/p(B)

$$
\begin{aligned}
I(m_j \to m_k) &= \sum p(m_o)\log\left(\frac{p(m_o)}{q(m_k)}\right) - \sum p(m_o)\log\left(\frac{p(m_o)}{q(m_j)}\right) \\
&= \sum p(m_o)\log\left(\frac{q(m_j)}{q(m_k)}\right)
\end{aligned}
$$

which is the weighted difference (weighted by the observed probabilities) of the difference between the logs of the two calculated probabilities

T and I are effect sizes, always positive, and I is more general than T

$$I(m_o \to m_j) = T(m_j) - T(m_o) = T(m_j)$$

$I(m_o \to m_j)$ is information lost in the model. (reference = top)
$I(m_j \to m_{ind})$ is information captured in a model. (reference = bottom)

We can only compare Transmission of models that are nested in the lattice of structure, i.e. they must be ancestors or descendents of each other.

OCCAM prints out information normalized by $T(m_{ind})$ so that information is between 0 and 1.

e.g. $\dfrac{I(m_j \rightarrow m_k)}{T(m_{ind})}$

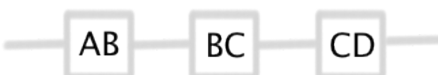
where $\dfrac{I(m_o \rightarrow m_{ind})}{T(m_{ind})} = 1$

L^2 is likelihood ratio, a measure of statistical significance of the effect size, (Kripendorff p. 87)

$L^2 = 1.3863\ n\ I$ For n = sample size

Krippendorff p. 44-45 information is additive for chain models

A chain model has a general structure that looks like a chain, with pairs of variables.



$I(m_0 \rightarrow m_{chain})$ = error in the chain model
$I(m_{chain} \rightarrow m_{ind})$ = information captured in the chain model

$I(m_o \rightarrow m_{ind}) = I(m_0 \rightarrow m_{chain}) + I(m_{chain} \rightarrow m_{ind})$
$T(A:B:C:\ldots) = T(AB:BC:\ldots) + T(A:B) + T(B:C) + \ldots$

OCCAM will let you search only for chain models.

3 – 4. Calculating q and IPF; Maximizing H Subject to Constraint

Algebraic Calculations

$$T(m_j) = \sum p(m_o) \log\left(\frac{p(m_o)}{p(m_j)}\right)$$

4 cases from simplest to most complex

1. $q(m_{ind})$: $q(A:B:C:\ldots) = p(A)\ p(B)\ p(C)\ldots$

2. disjoint; $q(AB{:}CDE{:}\ldots) = p(AB)\, p(CDE)\ldots$

3. overlap, no loops

$q(AB{:}BC) = p(AB)\, p(C|B) = p(AB)\, p(BC) / p(B)$

$q(AB{:}BC\ldots) = p(AB)\, p(C|B) = p(AB)\, p(BC) / p(B) \ldots$

$q(AB{:}BC{:}CDE) = p(AB)\, p(C|B)\, p(DE|C)$

4. loops By IPF (no algebraic solutions)

OCCAM always does IPF, which converges in one iteration when there are no loops.
Iterative Proportional Fitting

Consider the probability table for data AB

AB p table **df = 3**

	B_1	B_2	
A_1	**.1**	**.2**	.3
A_2	**.3**	.4	.7
	.4	.6	

Calculating the q table for A:B, **df = 2**

	B_1	B_2	
A_1	q_1	q_2	**.3**
A_2	q_3	q_4	.7
	.4	**.6**	

	B_1	B_2	
A_1	.12	.18	**.3**
A_2	.28	.42	.7
	.4	**.6**	

This method maximizes entropy subject to the constraints the margins, i.e. we want to maximize $H(q) = -\, q_1 \log q_1 - q_2 \log q_2 - q_3 \log q_3 - q_4 \log q_4$ [or $\Gamma(q_1, q_2, q_3, q_4)$], such that

$q_1 + q_2 = .3$

$q_3 + q_4 = .7$

$q_1 + q_3 = .4$

$q_2 + q_4 = .6$

$q_1 + q_2 + q_3 + q_4 = 1$

The last constraint is assumed, and the second and fourth are redundant. Therefore there are two constraints, and df = 2

To satisfy the two constraints,

$q_1 = .12$
$q_2 = .18$
$q_3 = .28$
$q_4 = .42$

Constraints could also be written as matrix-vector equation. For three constraints, α, β, γ

α: $q_1 + q_2 = .3$
β: $q_2 + q_4 = .6$
γ: $q_1 + q_2 + q_3 + q_4 = 1$

The matrix, M, is given by:

$$\begin{array}{c} \\ \alpha \\ \beta \\ \gamma \end{array} \begin{array}{c} \begin{array}{cccc} q_1 & q_2 & q_3 & q_4 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

and df = rank of M (# of rows) - 1

In IPF, entropy is maximized subject to the following equation:

$$M\check{q} = M\check{p}$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} .1 \\ .2 \\ .3 \\ .4 \end{bmatrix} = \begin{bmatrix} .3 \\ .6 \\ 1 \end{bmatrix}$$

To do IPF:
Start with uniform model.
Impose constraints one at a time.
If after posing all constraints, each constraint is still satisfied, IPF is done and the model does not contain loops. If some constraints are not satisfied, impose them again.

State-based model example where df = 2:

$q_1 + q_2 = .3$
$q_3 = .3$
$q_1 + q_2 + q_3 + q_4 = 1$

.1	.2	**.3**
.3	.4	.7
.4	.6	

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$