CULTURAL CONSENSUS THEORY: AGGREGATING EXPERT JUDGMENTS ABOUT TIES IN A SOCIAL NETWORKS

William H. Batchelder

Department of Cognitive Sciences

University of California Irvine

whbatche@uci.edu

SBP 09

AGENDA

I. What is Cultural Consensus Theory (CCT)?

II. A Social Network Consensus Model

III. Properties and Statistical Inference

IV. Adding Constraints to the Model

V. Relationship of CCT to Other Information Pooling Approaches

I. WHAT IS CULTURAL CONSENSUS THEORY?

- ❖ Cultural Consensus Theory (CCT) is an approach to information pooling (aggregation, data fusion).

- ❖ <u>Problem</u>

  - ✓ One has access to a few ‘informants’ who may share ‘cultural’ knowledge unknown to the researcher.

  - ✓ The researcher can construct questionnaire items but does not know the answers, if any, that best represent the shared knowledge.

  - ✓ Also the researcher does not know the ‘cultural competence’ and ‘response bias’ of each informant.

CCT APPLICATION AREAS

- ❖ Ethnographic studies in social and cultural anthropology, for example determining folk medical or religious beliefs
- ❖ Determining beliefs of a deviant group
- ❖ Inferring events that happened from eyewitness reports
- ❖ Discovering social relationships in a covert network
- ❖ Determining the syntax of an exotic language
- ❖ Aggregating probability estimates
- ❖ Data fusion from automated sources

TYPES OF QUESTIONS AND MODELS

- The questions represent shared knowledge of a group rather than individual preferences. For example:

  Do Hoosiers like corn?-- OK

  Do you like corn? -- Not OK

- Questions can be in various formats, e.g. true/false; multiple choice, matching, ranking, continuous scale.

- For each question format CCT specifies a cognitively valid parametric probability model of the informant. Models derive from **Psychometric Test Theory** (Item Response Theory) and **Signal Detection Theory-**- except the ‘correct answers’ are parameters in the model instead of being known apriori.

ORIGIN OF CCT

- CCT was created by A. Kimball Romney and myself in the 1980s. e.g. Batchelder and Romney (1986,1988,1989)

  Romney, Weller, Batchelder(1986), Romney, Batchelder, Weller (1987)- see reference list.

- It has become a major methodological tool in social, cultural and medical anthropology, and it has been used in other areas of the social and cognitive sciences.

THE NATURE OF THE DATA

- A questionnaire format is selected, *M* items are generated pertaining to the unknown shared knowledge, and responses to each question are collected from each of *N* informants.

- The data is represented in an *NxM* response profile matrix

$$\mathbf{X} = (X_{ik})_{NxM}$$

where X_{ik} is the response of informant *i* to question *k*.

GOALS OF CCT

- CCT uses the information in to: $\mathbf{X} = (X_{ik})_{NxM}$

1. Identify one or more latent ‘cultural groups’ of informants that share knowledge.

2. Decide if the statistical model used for (1) is valid, (a good enough approximation) and if valid, for each cultural group to:

2A. Estimate the consensus answers to the questions as well as their ‘difficulty.’

2B. Estimate the ‘competence’ and ‘response bias characteristics’ of each informant.

II. A SOCIAL NETWORK CONSENSUS MODEL

- A model is developed to aggregate responses from multiple expert sources (informants) about the ties in a social network.

- A simple social network is represented by a digraph on a fixed set of nodes.

- The nodes represent social entities, e.g. people, clubs, corporations, countries, etc.

- The directed ties from one node to another represent social connections, e.g. ‘socializes with’, ‘gives orders to’, ‘shares information with’, ‘trades with,’ etc.

VARIABLES AND PARAMETERS

- Each of *N* informants answers "yes" or "no" for each possible tie in the digraph of *M* nodes. Ties are treated as signals and the model is motivated by signal detection.

- Response Profile Data-- $\mathbf{X} = (X_{i,jk})_{NxMxM}\,,\; X_{i,jj} \equiv 0.$

$$X_{ik} = \{ \begin{matrix} 1 \text{ if informant } i \text{ answers "yes" to tie } jk \\ 0 \text{ if informant } i \text{ answers "no" to tie } jk \end{matrix}$$

- Answer Key-- $\mathbf{Z}^* = (Z^*_{jk})_{MxM}\,, Z^*_{jj} \equiv 0$

$$Z^*_{jk} = \{ \begin{matrix} 1 \text{ if there really is a tie from } j \text{ to } k \\ 0 \text{ if there is not a tie from } j \text{ to } k \end{matrix}$$

MORE VARIABLES AND PARAMETERS

- Performance Profile Data-- $\mathbf{Y} = (Y_{i,\,jk})_{NxMxM}$

$$Y_{i,\,jk} = \{ \begin{matrix} 1 \text{ if informant } i \text{ correct on tie } jk \\ 0 \text{ if informant } i \text{ incorrect on tie } jk \end{matrix}$$

- Hit Rates -- $\mathbf{H} = (H_{i,jk})_{NxMxM}$

$$\forall\ j \neq k,\ H_{i,jk} \in (0,1),\ \ H_{i,jj} \equiv 1$$

- False Alarm Rates -- $\mathbf{F} = (F_{i,jk})_{NxMxM}$

$$\forall j \neq k,\ F_{i,jk} \in (0, H_{i,jk}),\ \ F_{i,jj} \equiv 0$$

AXIOMS FOR THE DIGRAPH MODEL

AXIOM 1: (**Unitary Consensus**) There is a single answer key $\mathbf{Z}^* = (Z^*_{jk})_{MxM}$ applicable to all informants.

AXIOM 2: (**Conditional Independence**) The response profile matrix satisfies conditional independence:

$$\Pr[\mathbf{X} = (x_{i,jk}) \big| \mathbf{Z}^* = (z^*_{jk}), \mathbf{H} = (H_{i,jk}), \mathbf{F} = (F_{i,jk})]$$

$$= \prod_{i=1}^{M} \prod_{j=1}^{N} \prod_{\substack{k=1 \\ k \neq j}}^{N} \Pr(X_{i,jk} = x_{i,jk} \big| Z^*_{jk} = z^*_{jk}, H_{i,jk}, F_{i,jk})$$

for all possible realizations $(x_{i,jk}),\ (z^*_{jk}), (H_{i,jk}),\ (F_{i,jk})$

MORE AXIOMS

AXIOM 3:(Signal Detection)

$$\Pr(X_{i,jk} = 1 \mid Z^*_{jk} = z^*_{jk}, H_{i,jk}, F_{i,jk}) =$$

$$\left\{ \begin{array}{l} H_{i,jk} \text{ if } z^*_{jk} = 1 \\ F_{i,jk} \text{ if } z^*_{jk} = 0 \end{array} \right.$$

- ❖ The model at this point is way over parameterized. There are 2*NM*(*M*-1) hit and false alarm parameters, *M*(*M*-1) answer key parameters, and only *NM*(*M*-1) bits in the response profile matrix **X**.

REDUCING PARAMETERS

AXIOM 4a: (**Out-arc Homogeneity**)

$$\forall 1 \le j, k \le M, \; j \ne k, \; H_{i,jk} = H_{i,j\bullet}, \; F_{i,jk} = F_{i,j\bullet}$$

- ❖ Axiom 4a assumes hits and false alarm rates for ties from a given node (out-arcs) are constant for each informant, but they can vary from informant to informant and within an informant from node to node.

AXIOM 4b: (**In-arc Homogeneity**)

$$\forall 1 \le j, k \le M, \; j \ne k, \; H_{i,jk} = H_{i,\bullet k}, \; F_{i,jk} = F_{i,\bullet k}$$

MORE PARAMETER REDUCTION

OBSERVATION 1. Axioms 4a and Axiom 4b together imply homogeneous informants

$$H_{i,jk} = H_{i,\bullet\bullet},\ F_{i,jk} = F_{i,\bullet\bullet}$$

- Choosing one of Axioms 4a and 4b allows informants to have differential competency at different nodes of the digraph. Hereafter we analyze Axioms 1,2,3,4a.

- There are 2*NM* hit and false alarm rates, *M*(*M*-1) answer key parameters, and *NM*(*M*-1) data bits.

III. PROPERTIES AND STATISTICAL INFERENCE

OBSERVATION 2. Assume Axioms 1,2,3,4a. Then if N>2, M>3 there are more data bits than parameters.

- Assuming this model, we can analyze the data node by node and determine the consensus out-arcs.

- For a given node *j* define the random matrix $\mathbf{R}_j = (X_{i,jk})$. We use the data realizations of **R***j* to infer things about the parameters --

$$\mathbf{Z}_j^* = (Z_{jk}^*)_{1xM}\,;\, \mathbf{H}_j = (\,H_{i,j\bullet})_{1xN}\,,\, \mathbf{F}_j = (F_{i,j\bullet})_{1xM}$$

LIKELIHOOD FUNCTION

OBSERVATION 3. Given Axioms 1,2,3,4a and **R**$_j$, the likelihood function is given by

$$L(\mathbf{H}_j, \mathbf{F}_j, \mathbf{Z}_j^* | \mathbf{R}_j) =$$

$$\prod_{i=1}^{N} \prod_{\substack{k=1 \\ k \neq j}}^{M} \left[\frac{H_{i,j\bullet}(1-F_{i,j\bullet})}{F_{i,j\bullet}(1-H_{i,j\bullet})} \right]^{x_{i,jk} z_{jk}} \left[\frac{1-H_{i,j\bullet}}{1-F_{i,j\bullet}} \right]^{z_{jk}} \left[\frac{F_{i,j\bullet}}{1-F_{i,j\bullet}} \right]^{x_{i,jk}} \left[1-F_{i,j\bullet}\right]$$

With parameter spaces $\forall i, \forall j; 0 < F_{i,j\bullet} \leq H_{i,j\bullet} < 1$ and

$$\mathbf{Z}_{jk}^* \in \{0,1\}^{M-1}$$

M.L.E.s GIVEN HIT AND FALSE ALARM RATES

OBSERVATION 4. Given realizations of $<\mathbf{H}_j, \mathbf{F}_j>$,and a uniform prior, the posterior distribution of $\mathbf{Z}_j^*$ is maximized by

$$\hat{Z}_{jk}^* = 1 \text{ iff } \sum_{i=1}^{N} x_{i,jk} \ln\left[\frac{H_{i,j\bullet}(1-F_{i,j\bullet})}{F_{i,j\bullet}(1-H_{i,j\bullet})}\right] \geq \sum_{i=1}^{N} \ln\left(\frac{1-F_{i,j\bullet}}{1-H_{i,j\bullet}}\right)$$

- ❖ Observation 5 shows that the M.L.E. weights each informant’s response by a type of log odds. This is where the model can out perform ‘majority rule.’
- ❖ Given $\mathbf{Z}_j^*$ formula from signal detection give $<\mathbf{H}_j, \mathbf{F}_j>$

INFORMANT BY INFORMANT CORRELATIONS

- We can define a correlation coefficient $r_{u,v}(j)$ between pairs of informants *u* and *v* across out-arcs from *j* from the two by two table

		$X_{v,j}$	
		1	0
$X_{u,j}$	1	A_{11}	A_{10}
	0	A_{01}	A_{00}

where A_{11} is the number out of M-1 possible out-arcs from node *j* that both informants *u* and *v* assess as present, etc. Let $\rho_{u,v}(j) = E(r_{u,v}(j) \vert \mathbf{Z}^*_j = (Z^*_{jk})_{1xM})$

A SPEARMAN PROPERTY OF THE GCM

- **OBSERVATION 5**. Suppose the Digraph Model holds. Then $\forall 1 \le u, v \le N, u \ne v, \ \forall 1 \le j \le M,$

$$\rho_{u,v}(j) = \rho_{u,\mathbf{Z}_j^*} \rho_{v,\mathbf{Z}_j^*}$$

$\rho_{u,\mathbf{Z}_j^*}$ Is interpreted as the correlation of an informant with the (latent) truth.

- This relationship says that the correlation across the out-arcs of any node between two informants is the product of their separate correlations with the answer key. This is a form of Spearman’s famous law of tetrads--

$$\rho_{i,j}\rho_{k,l} = \rho_{i,l}\rho_{k,j}$$

A USEFUL MODEL CHECK

- Observation 6 suggests a useful model check.

- Let $\mathbf{C}_j = (c_{uj;vj})_{NxN}$ be a matrix of empirical informant by informant covariances from their out-arc responses to node *j*. Then **C**$_j$ has a one factor structure in the sense that the off-diagonal terms approximately satisfy $\mathbf{C}_j \cong \mathbf{V}_j^T \cdot \mathbf{V}_j$ where

$$\mathbf{V}_j = (v_{i,j})_{1xN} \qquad v_{i,j} \propto H_{i,j} - F_{i,j}$$

- This one-factor property is easily checked. The components of **V**$_j$ are estimates of the ‘competence’ of the informants. They should be positive, and additional factors of **C**$_j$ should be ‘noise’

ESTIMATING THE MODEL

- The main objective of CCT is to estimate the consensus answers. In the Digraph Model they are discrete so combinatorial optimization is required. We have used simulated annealing (Batchelder, Kumbasar, Boyd, 1997, J*. Math. Sociology*).

- More recently we have adopted a Bayesian formulation and used MCMC methods to estimate a closely related CCT model (Karabatsos & Batchelder, 2003, *Psychometrika*).

- There is much more work to be done to develop a general Bayesian inference package for this and other CCT models.

INTERPRETING PARAMETERS

- An advantage of our formulation is that we estimated the consensus out-arcs from hits and false alarm parameters rather than the parameters of a particular signal detection model, e.g. TSD, 2HTM, 1HTM.

- Clearly to interpret the hits and false alarms they need to be reparameterized into the parameters of a standard signal detection model.

- We have done this for assessing out-arcs in a friendship network defined over workers in a hardware production firm.

  Batchelder, Kumbasar, Boyd (1997, *J. Math. Sociology*).

IV. Adding Constraints to the Model

- ❖ Perhaps the weakest feature of the model is Axiom 2, local independence (LI) --

$$\Pr[\mathbf{X} = (x_{i,jk}) \big| \mathbf{Z}^* = (z^*_{jk}), \mathbf{H} = (H_{i,jk}), \mathbf{F} = (F_{i,jk})]$$

$$= \prod_{i=1}^{M} \prod_{j=1}^{N} \prod_{\substack{k=1 \\ k \neq j}}^{N} \Pr(X_{i,jk} = x_{i,jk} \big| Z^*_{jk} = z^*_{jk}, H_{i,jk}, F_{i,jk})$$

- ❖ It is a strong form of LI because responses to ties are conditionally independent given parameters only concerning that tie.

- ❖ Most social networks have structural features that could influence response probability to specific ties.

MORE ON LOCAL INDEPENDENCE

- Violations of local independence are well studied in Item Response Theory (IRT). There is robustness but—

- I am working on an approach that maintains conditional independence and puts more of the Digraph into the condition, for example adding elements of **Z*** that influence a tie (*j,k*).

- This is the approach taken in Markov random field approach to graph models (e.g. Wasserman and Pattison(1996, Psychometrika). Here one could replace Z^*_{jk} by

$$A^*_{jk} = \{(u,v) \big| \{u,v\} \cap \{i, j\} \neq \emptyset\}$$

ADDING GRAPH CONSTRAINTS

- I am working on an approach to add constraints such as symmetry, transitivity, balance, etc. to the consensus digraph.

- Informant’s are free to respond but the estimated consensus digraph must satisfy some relational axioms.

- The barrier is that one must solve a combinatorial optimization problem by searching only digraphs that satisfy the axioms. There are $2^{M(M-1)}$ digraphs.

- One idea is to use ‘well gradedness’ (e.g. Doignon & Falmagne, 1997, Discrete Math.) and simulated annealing.

V. RELATIONSHIP OF CCT TO OTHER APPROACHES TO INFORMATION POOLING

- In the computational sciences there are a number of useful approaches to information pooling (aggregation, fusion).

- CCT specifies validated cognitive or psychometric models of the informant that include both competence and response bias parameters augmented with ‘consensus answer' parameters.

- Informants must answer the same set of questions, but no prior interaction or technical knowledge is required.

- Parameters are estimated from the response profile data alone. No exogenous knowledge and no axiomatic information combining rules are imposed.

SELECTED REFERENCES

- Batchelder, W.H. (in press) . Cultural consensus theory: Aggregating expert judgments about ties in a social network. This Conference Volume.

- Batchelder, W.H., Kumbasar, E., and Boyd, J.P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology*, **22**, 29-58.

- Batchelder, W.H. and Romney, A.K.(1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman and G. Owen (eds.) *Information Pooling and group decision Making: Proceedings of the Second University of California Irvine Conference on Political Economy* (pp.103-112).Greenwich, Conn.: JAI Press .

- Batchelder, W.H. and Romney, A.K. (1988). Test theory without an answer key. *Psychometrika*, **53**, 71-92.

- Batchelder, W.H. and Romney, A.K. (1989). New results in test theory without an answer key. In E.E. Roskam (Ed.) *Mathematical Psychology in Progress* (pp.229-248).Heidelberg: Springer-Verlag.

- Karabatsos, G., and Batchelder, W.H. (2003). Markov chain estimation theory methods for test theory without an answer key. *Psychometrika*, **68**, 373-389.

- Romney, A.K., and Batchelder, W.H. (1999). Cultural consensus theory. In R. Wilson and F. Keil (eds.). The MIT Encyclopedia of the Cognitive Sciences. Cambridge, MA: The MIT Press.

- Romney, A.K., Batchelder, W.H., Weller, S.C. (1987). Recent applications of cultural consensus theory *American Behavioral Sciences*. **31**, 129-141.
- Romney, A.K., Weller, S.A., and Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, **88**, 313-338.