Cultural Consensus Theory for the Evaluation of Patients' Behavior in Psychiatric Detention Centers

Don van den Bergh*[1], wie nog meer?[1], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam

Abstract

In many psychiatric detention centers, patients' mental health is monitored at regular intervals. Typically, clinicians score patients using a Likert scale on multiple criteria including hostility. Having an overview of patients' scores benefits staff members in at least three ways. First, the scores may help adjust treatment to the individual patient; second, the change in scores over time allow an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent. Practical importance notwithstanding, current practices for the analysis of mental health scores are suboptimal: evaluations from different clinicians are averaged (as if the Likert scale were linear and the clinicians identical), and patients are analyzed in isolation (as if they were independent). Uncertainty estimates of the resulting score are often ignored. Here we outline a quantitative program for the analysis of mental health scores using cultural consensus theory (CCT; Anders & Batchelder, 2015). CCT models take into account the ordinal nature of the Likert scale, the individual differences among clinicians, and the possible commonalities between patients. In a simulation, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and, as a more reasonable alternative, against often-used machine learning toolboxes. In addition, we outline the substantive conclusions afforded by application of the CCT model. We end with recommendations for clinical practitioners who wish to apply CCT in their own work.

*Correspondence concerning this article should be addressed to:
Don van den Bergh
University of Amsterdam, Department of Psychological Methods
Postbus 15906, 1001 NK Amsterdam, The Netherlands
E-Mail should be sent to: donvdbergh@hotmail.com.

Psychiatric detention centers monitor the mental health of their patients at regular intervals. A clinician, psychiatrist, or other staff member, henceforth a *rater*, scores a patient on multiple criteria. For example, a rater evaluates a patient's behavior on a variety of criteria that relate to aggressiveness. Next, these ratings of patients' mental health are used for a variety of purposes. For instance, the scores may to help adjust treatment to individual patients; second, the change in scores over time allows for an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent. Moreover, these ratings are key to a quantitative approach of describing and predicting patients' behavior.

Current practices for aggregating the scores are suboptimal. Evaluations from different raters are averaged, as if they are exchangeable. For example, personal communication with staff of a psychiatric detention center suggested that clinicians are more lenient in their ratings than psychiatrist, but this information is not used to weigh their ratings. Furthermore, different patients are analyzed in isolation, as if they are independent. Any information regarding a patient's criminal offense is not accounted for in a model-based manner. In addition, any uncertainty estimates of the resulting score are usually ignored.

Don: lookup quote

An appropriate model for these data captures individual differences among the patients, raters, and items. Cultural Consensus Theory (CCT) is an ideal starting point for such a model, as CCT is designed to pool information from different raters and items (Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988; Batchelder & Anders, 2012). The CCT model can be applied to both continuous as well as ordinal describe raters and items using a hierarchical structure, which allows

Here we outline a quantitative program for the analysis of mental health scores using CCT. First, a CCT model for ordinal data is introduced (Anders & Batchelder, 2015). Afterwards, this model is expanded step by step, in order to include more characteristics of the data. In a simulation, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and, as a more reasonable alternative, against often-used machine learning toolboxes such as Random Forest (Breiman, 2001) and Boosted Regression Trees (Friedman, 2002). We showcase the substantive conclusions obtained from applying the CCT model and conclude the paper with recommendations for clinical practitioners who wish to apply CCT in their work.

Cultural Consensus Theory

Don: Is het leuk om een korte historische context the geven? Persoonlijk boeit dat mij meestal niet maar sommige lezers zouden het leuk kunnen vinden.

Historically, CCT used in context of questionnaires where the 'true' answers are unknown and estimated from the data. Examples: "unknown answer key",

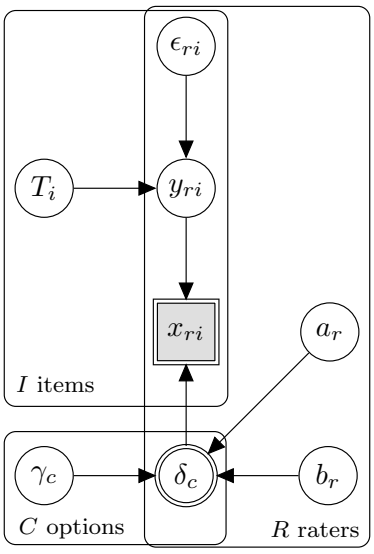

Figure 1: Graphical model corresponding to the CCT model for a single patient.

political questionnaire.

As a starting point, consider the Cultural Consensus Model for ordinal data as introduced by Anders and Batchelder (2015). This model captures differences among raters and items, and may be viewed as the simplest model for a single patient. Here, we use a model for ordinal ratings, but CCT can be applied to continuous data as well (e.g., the LTM; Batchelder & Anders, 2012). A graphical model is shown in Figure 1.

To formally introduce the CCT model for the ordinal case, the rating given by rater r on item i is denoted x_{ri}, and can take on discrete values from 1 through C. The realization of x_{ir} is assumed to follow an ordered logistic distribution* with location y_{ir} and thresholds δ_{rc}:

$$P(x_{ir} \mid y_{ir}, \delta_r) = \begin{cases} 1 - \text{logit}^{-1}\left((y_{ir} - \delta_{r1}\right) & \text{if } x_{ir} = 1, \\ \text{logit}^{-1}\left(y_{ir} - \delta_{r,c-1}\right) - \text{logit}^{-1}\left(y_{ir} - \delta_{r,c}\right) & \text{if } 1 < x_{ir} < C, \\ \text{logit}^{-1}\left(y_{ir} - \delta_{r,C-1}\right) & \text{if } x_{ir} = C. \end{cases}$$

The ordered logistic distribution uses $C - 1$ thresholds to assign each of the C outcomes a probability. This probability is equal to the area between subsequent thresholds for a logistic distribution, as illustrated in Figure 2.

Don: Als ik zo naar het model kijk vraag ik me af waarom we de bias in de latent appraisal nodig hebben, en of deze niet al opgevangen wordt door de rater specifieke thresholds.

*The choice for an ordered logistic distribution is arbitrary and an ordered probit distribution could also be used.

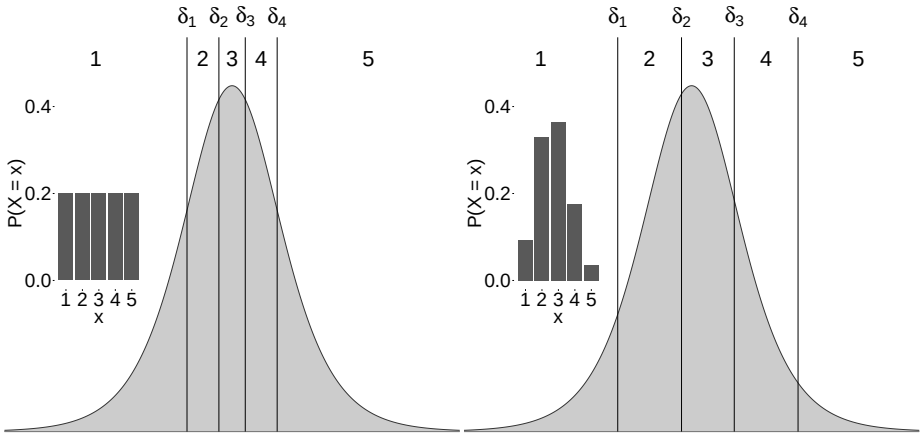

Figure 2: Ordered Logistic Distribution for $y_{ir} = 0$ and varying thresholds. The implied probability distribution over categories is shown inside each panel. In the left panel, the thresholds are unbiased, $\alpha_r = 1$, $\beta_r = 0$. As a consequence, the distribution over outcomes is uniform. In the right panel, the thresholds are shifted right and the scale increased slightly, $\alpha_r = 2$, $\beta_r = 0.5$. The distribution over outcomes has most mass on outcomes 2 and 3.

The main purpose of the ordered logistic distribution is to translate each observed rating into some latent appraisal score y_{ir} which can subsequently be linked to patient, rater, and item characteristics. The model contains two components that can capture rater bias. First, the appraisal of a rater for a given item can be biased. Therefore the appraisal is the sum of two components, the true score for an item T_i and the bias of the rater ϵ_r. This captures that ..., Second, the thresholds are allowed to differ across raters. An initial guess for the thresholds is $\text{logit}^{-1}(c/C)$. This yields a set of thresholds such that if the latent appraisal is 0 then $P(x_{ir})$ is uniform.

Starting from an initial guess that the thresholds are distributed uniformly over the logistic distribution (here uniformly means, i.e., the),

Rather than modeling each of the $C-1$ thresholds individually, the thresholds are modeled as a deviation from an unbiased set of thresholds. The thresholds are obtained by shifting and rescaling an unbiased threshold placement to estimate a potentially large number of thresholds instead of needing a parameter per threshold (Fox & Tversky, 1995; Gonzalez & Wu, 1999).

vary across raters. Rather than modeling each of the $C-1$ thresholds individually, we adopt the approach to

The thresholds are obtained by shifting and rescaling an unbiased threshold placement to estimate a potentially large number of thresholds instead of needing a parameter per threshold (Fox & Tversky, 1995; Gonzalez & Wu, 1999).

The appraisals consists of the latent truth T_i for each item and the bias of each rater ϵ_r. The thresholds are obtained by shifting and rescaleing

to estimate a potentially large number of thresholds instead of needing a parameter per threshold (Fox & Tversky, 1995; Gonzalez & Wu, 1999).

personen/ raters zijn gebiased. Kan niet alleen eigenschappen van de gedetineerden halen maar ook die van de raters.

Covariaat voor e.g., hulpverleners versus psychiaters. Meerdere groepen raters (fixed effect).

Hierachisch niveau over patienten.

Covariaat voor groepen gedetineerden (fixed effect misdrijf). Ofwel order restrictie voor deze groepen.

Missing values?

Hoe combineren we verschillende schalen van items?

kaart van hoe ontwikkelt zich dit over de tijd (zie figuur in proposal)

TODO: change model in NWO proposal to mimic SDT approach.

Data is simulated using R (R Core Team, 2019) and the posterior distributions are sampled from using Stan (Carpenter et al., 2017).

Simulation Results

Discussion

A common application of CCT is testing for multiple consensus truths.

Limitations

Aanbeveling voor de praktijk

Bijhouden van evaluaties (scores + rater), liefst met hoge frequentie. Reden waarom iemand opgesloten is (reason of incarceration). zo min mogelijk missing not at random. "handig": training in invullen om verschillen tussen raters te minimalizeren. (V)AR component?

References

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The quarterly journal of economics*, *110*(3), 585–603.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, *38*(1), 129–166.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American anthropologist*, *88*(2), 313–338.