Todo list

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

Don van den Bergh*[1], Stefan Bogaerts[2], Marinus Spreen[3], Rob Flohr[4], Joachim Vandekerckhove[5], Mijke Rhemtulla[6], William Batchelder[5], and Eric-Jan Wagenmakers[1]

[1]University of Amsterdam
[2]University of Tilburg
[3]Mesdag Clinic
[4]Stenden University of Applied Sciences
[5]University of California Irvine
[6]??

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

*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.

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.

Psychiatric detention centers monitor the mental health of their patients at regular intervals. A clinician, psychiatrist, or another 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 multiple purposes. For instance, the scores may 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 to monitoring and forecasting 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 the staff of a psychiatric detention center suggested that clinicians are more lenient in their ratings than psychiatrists, 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: quote opzoeken over non exchangeability van raters/ items.

The goal of a model appropriate for these data should be to disentangle the patient, rater, and item-specific components in the data. Such a model would benefit decision making because it separates these three different sources of variances and thus decision can be based on estimates without rater of measurement error. For example, such a model should yield estimates for a patient's aggressiveness, while accounting for rater bias, item-specific measurement error, and a patient's criminal offense. An ideal modeling framework for this is Bayesian hierarchical modeling (Shiffrin, Lee, Kim, & Wagenmakers, 2008). In particular, Cultural Consensus Theory (CCT), which is designed to pool information from different raters and items, could be seen as a model that accomplishes exactly this – but for a single patient. (Romney, Weller, & Batchelder, 1986; Batchelder & Romney, 1988; Batchelder & Anders, 2012).

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). Afterward, this model is expanded step by step, in order to include more characteristics of the data. We demonstrate how CCT could be used to monitor patients progress over time. 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

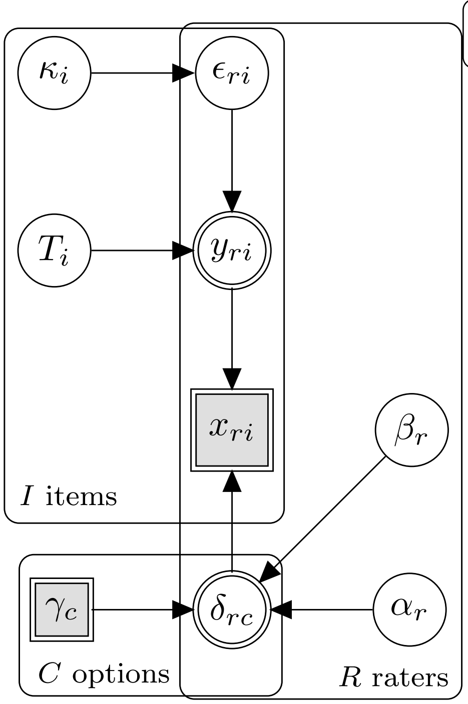
The next sections introduce Cultural Consensus Theory (CCT). First, a brief introduction to CCT is given. Afterward, the CCT model developed in (Anders & Batchelder, 2015) is introduced, which serves as the simplest model for a single patient. Subsequently, this model is expanded in three steps. First, the model is expanded to describe multiple patients. Next, latent constructs are added to the model. Finally, the model is modified to include patient and rater characteristics.

Cultural Consensus Theory, also known as "test theory without an answer key" (Batchelder & Romney, 1988), is a statistical tool that attempts to retrieve the "truth" for an item by examining the consensus among the responses. The model accounts for varying difficulty in items and varying competence among respondents. In addition, the model can be expanded to allow for multiple consensus truths, where there may be multiple unknown truths that vary across subgroups of respondents (Anders & Batchelder, 2012). For example, given a political questionnaire, there is no objectively correct answer key. Instead, one could administer the questionnaire to left-oriented respondents and use CCT to find out what the consensus is among left oriented respondents. Rather than selecting left-oriented respondents as a target population, the questionnaire could also be given to people at random. That way, the different consensuses (e.g., left, right, center, etc.) could be derived from the data. The property of CCT models to estimate the consensus truth for an item from the data is ideal for psychiatric data, where a patient's true state is unknown and a consensus from the raters is desired.

The next sections formally introduce and expand the CCT models used in this paper. As a starting point, consider the Latent Truth Rater Model cultural consensus model for ordinal data (LTRM) as introduced by Anders and Batchelder (2015). We use a model for ordinal data as ratings are generally given on a Likert scale, but CCT can be applied to continuous data as well (e.g., the LTM; Batchelder & Anders, 2012). Figure 1 shows a graphical model of the LTRM. The LTRM captures differences among raters and items and may be viewed as the simplest model for a single patient.

Don: Definieer Ordered Logistic als een kansverdeling met locatie, scale en thresholds als parameters.

The rating given by rater r on item i is denoted x_{ri} and takes on discrete values from 1 through C. Anders and Batchelder (2015) formalize the core ideas of the LTRM with 6 axioms. There is a latent, shared cultural truth among the raters, which is captured by the item location parameters T_i (Axiom 1). Since raters are not perfect measurement instruments, they infer a noisy version of the



Don: Hierarchische componenten missen

$$x_{ri} \leftarrow \begin{cases} 1 \text{ if } y_{ri} \leq \delta_{r1} \\ c \text{ if } \delta_{r,c-1} < y_{ri} \leq \delta_{rc} \\ C \text{ if } y_{ri} > \delta_{r,C-1} \end{cases}$$

$$y_{ri} \leftarrow T_i + \epsilon_{ri}$$
$$\gamma_c \leftarrow \text{logit}\,(c/C)$$
$$\delta_{rc} \leftarrow a_r\gamma_c + b_r$$
$$\sigma^{\epsilon}_{ri} \leftarrow \kappa_r/\zeta_r$$
$$\sigma^{\epsilon}_{ri} \leftarrow \kappa_r/\zeta_r$$
$$\epsilon_{ri} \sim \text{Normal}\,(0,\, \sigma^{\epsilon}_{ri})$$
$$T_i \sim \text{Normal}\,(0,\, 10)$$
$$\kappa_i \sim \text{Gamma}\,(0.01,\, 0.01)$$
$$a_r \sim \text{Gamma}\,(0.01,\, 0.01)$$
$$b_r \sim \text{Normal}\,(0,\, 10)$$

Figure 1: Graphical model corresponding to the LTRM; a CCT model for a single patient. Gray nodes are observed whereas white nodes are unobserved parameters. Square nodes are discrete and circular nodes are continuous. A double border implies that a node is ???.

cultural truth for each item, called a latent appraisal and defined as $y_{ri} = T_i + \epsilon_{ri}$, where $\epsilon_{ri} \sim \text{Normal}\,(0,\, \sigma^{\epsilon}_{ri})$ (Axiom 2). The standard deviation of the appraisal error σ^{ϵ}_{ri} reflects that raters differ in their precision, captured by ζ_r, and that items differ in their difficulty, captured by λ_r. These two components define the standard deviation, $\sigma^{\epsilon}_{ri} = \kappa_r/\zeta_r$ (Axiom 3). Latent appraisals are assumed to be conditionally independent given the latent truth T_i and the appraisal error ϵ_{ri} (Axiom 4). So far, the axioms describe a latent process that underlies each observation. To translate these latent appraisals to observed categorical responses, it is assumed that there exist $C - 1$ ordered thresholds δ_{rc}, such that each x_{ri} is generated in the following way (Axiom 5):

$$x_{ri} = \begin{cases} 1 & \text{if } y_{ri} \leq \delta_{r1} \\ c & \text{if } \delta_{r,c-1} < y_{ri} \leq \delta_{rc} \\ C & \text{if } y_{ri} > \delta_{r,C-1} \end{cases}$$

In practice, the generating process of x_{ri} is probabilistic and described with an ordered logistic distribution[1], which gives:

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

The thresholds δ_{rc} are not fixed across raters but accommodate the response biases of the raters. Anders and Batchelder (2015) do so by estimating $C - 1$ ordered thresholds γ and defining $\delta_{rc} = a_c\gamma_c + b_c$ (Axiom 6). This translation of thresholds is called the Linear in Log Odds function and is a useful tool for capturing bias in probability estimation (Fox & Tversky, 1995; Gonzalez & Wu, 1999; Anders & Batchelder, 2015).

A technical difficulty of estimating $C-1$ is that the number of thresholds γ_c increases with the number of response options. This introduces a large number of parameters that can be difficult to estimate, in particular when some response options are not observed (i.e., when there are ceiling or floor effects). In addition, the model is only identified if the sum of thresholds is zero ($\sum_{c=1}^{C} \gamma_c = 0$; otherwise adding a constant to T_i and γ_c yields an identical likelihood). Rather than modeling each threshold individually, we model the thresholds as deviances from an initial guess, $\gamma_c = \operatorname{logit}(c/C)$. This yields a set of thresholds such that if the latent appraisal is 0 then $P(x_{ri})$ is uniform. Response biases are incorporated in the same manner: $\delta_{rc} = a_r \operatorname{logit}(c/C) + b_r$. This simplification can still capture a wide variety of data sets (Selker, van den Bergh, Criss, & Wagenmakers, 2019).

Figure 2 provides an intuition for how the ordered logistic distribution can model different data sets. In the left panel, there is no response bias, $a_r = 1$ and $b_r = 0$, which yields a uniform distribution over the outcomes. In the right panel, an increase in response scale, $a_r = 2$, concentrates the responses around the middle outcomes. An increase in response shift, $b_r = .5$ moves the average response towards lower categories. The latent appraisal is the same in both panels.

The first extension is to allow the model to describe multiple patients. This can be achieved by allowing all item specific parameters to vary across patients (T_{ip}, κ_{ip}, which induces $\epsilon_{rip}, y_{rip}, x_{rip}$). This implies that the latent cultural truth of an item and the difficulty of an item vary across patients. A second extension is that we are often not just interested in the latent truth of a single item, but also in a construct that is measured by multiple items. This is essentially a factor model, where a latent construct (e.g., aggressiveness) is measured through multiple items. For this particular application, however, the items may contain rater bias which is simultaneously estimated and corrected for. Formally, there are L constructs captured by latent variables which are indicated by some items. The measurement model, i.e., which items load on what construct, is assumed to be known.

[1] The choice for an ordered logistic distribution is arbitrary and an ordered probit distribution could also be used, as was done by Anders and Batchelder (2015).

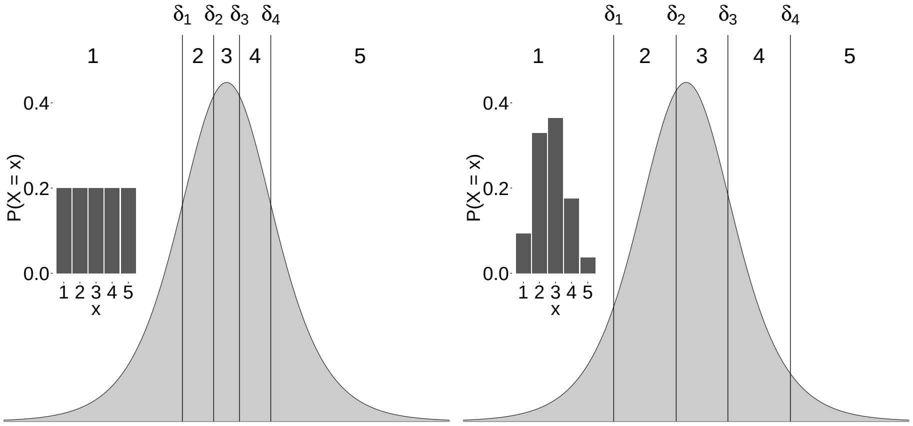


Figure 2: Ordered Logistic Distribution for $y_{ri} = 0$ and varying thresholds. The implied probability distribution over categories is shown inside each panel. In the left panel, there is no response bias, $\alpha_r = 1$, $\beta_r = 0$. As a consequence, the distribution over outcomes is uniform. In the right panel, the thresholds are shifted right and the scale increased slightly, $\alpha_r = 2$, $\beta_r = 0.5$. The distribution over outcomes is peaked on 2 and 3.

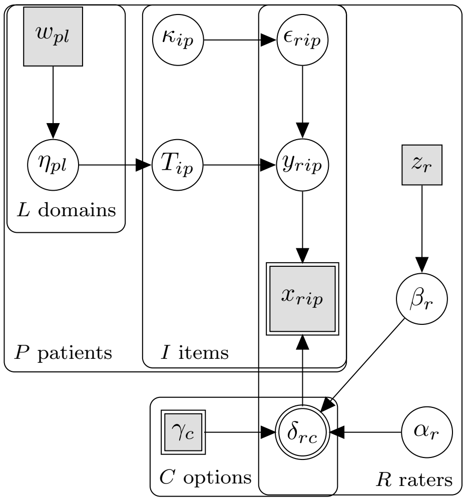
A third extension is to add additional information about raters and patients to the model. This would facilitate the model to capture that, for instance, child molesters are less aggressive than murderers. Such patient and raters characteristics could be perceived as covariates and their influence could be captured through a regression. For instance, the crime of a patient can be regressed on each latent variable, allowing the model to capture that child molesters and murderers differ in aggression, but not in say ... . Rater characteristics could be regressed directly on the latent appraisal. Adding these three components to the model results in the graphical model shown in Figure 3.

Don: ik heb eigenlijk geen idee wat ze allemaal meten

Simulation Results

The sections below illustrate the LTRM in a variety of scenarios. First, we demonstrate that the parameters of the LTRM can be accurately retrieved. Afterward, we show how the LTRM could monitor the progress of a patient across measurement occasions. Finally, we compare the predictive performance of the LTRM to the unweighted mean of the observations and two machine learning toolboxes.

In all simulations, data were simulated using R (R Core Team, 2019) and the posterior distributions were approximated with variational inference using Stan (Carpenter et al., 2017). We opted to use variational inference over traditional Markov chain Monte Carlo because it was computationally fast while providing



$$\begin{aligned}
T_{ip} &\sim \text{Normal}\left(\lambda_{il}\eta_{lp},\ 1\right)\\
\eta_{lp} &\sim \text{Normal}\left(w_{pl},\ 10\right)\\
w_{pl} &\sim \text{Normal}\left(0,\ 10\right)\\
b_r &\sim \text{Normal}\left(z_r,\ 10\right)\\
z_r &\sim \text{Normal}\left(0,\ 10\right)
\end{aligned}$$

Figure 3: Graphical model corresponding to the CCT model for multiple patients. Rater characteristics are captured by z_r and patient characteristics are captured by w_p.

similar results in terms of parameter retrieval and model predictions. Random Forest and Boosting analyses were done using the R packages `ranger` and `gbm` respectively (Wright & Ziegler, 2017; Greenwell, Boehmke, Cunningham, & Developers, 2019). The tuning parameters of the machine learning methods were optimized using the R package `caret` (Kuhn et al., 2008). R files and Stan models are available in the online appendix at *osflink*.

Parameter Retrieval

Before interpreting a model, it is key to assess if its parameters can be accurately retrieved. For this purpose, we simulated a data set of 50 patients, 30 raters, 20 items, and 5 answer categories. A patient-specific covariate containing 5 levels was added to mimic the effect of a patient's criminal offense. Similarly, three different categories of raters were assumed, which was also captured as a covariate. Parameter retrieval is shown in Figure 4, which graphs the values used to simulate data with against the posterior mean for each parameter.

All parameters are retrieved adequately. An exception is κ, which estimates appear more variable as the true value κ increases. However, the spread in the posterior means of κ is similar to that in Figure 6 of Anders and Batchelder (2015). Also, the posterior means of μ_β are approximately equal to the sample means of β_r split by the rater specific covariates. These sample means differ somewhat from the true means since there are 10 observations per mean.

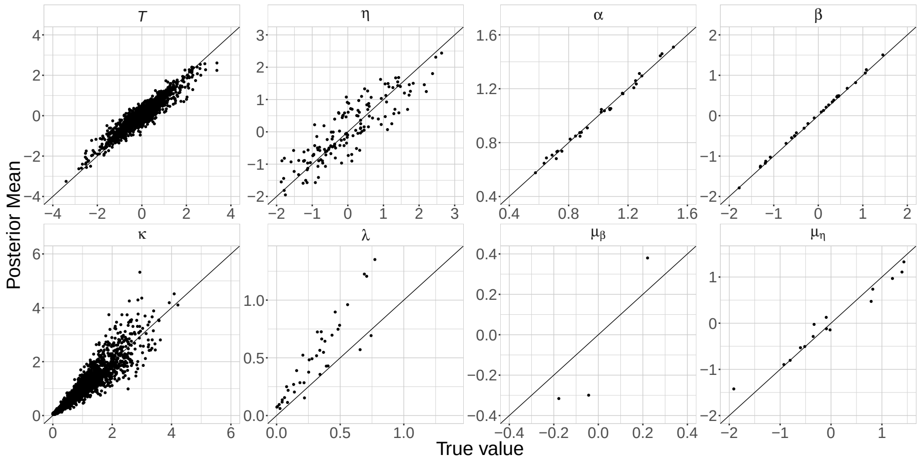


Figure 4: True value used for data simulation (x-axis) and posterior mean of that parameter (y-axis), for all parameters of the LTRM. Above each panel is indicated which parameter is shown.

Progress Monitoring

Ideally, patients are monitored over a period of time and data from multiple measurement occasions is obtained. Next, the LTRM is applied to analyze the data. However, the LTRM should not be applied to data from individual measurement occasions, because an identification restriction of the LTRM is that the mean of each latent construct is zero. Instead, the LTRM can be easily expanded with an intercept that varies across patients, latent constructs, and measurement occasions. As a result, the progress for a single patient on one construct could be visualized as in Figure 5.

Predictive Performance

Here, the predictive performance of the LTRM is compared to that of the mean and, as a more informative comparison, to Random Forest and Boosted Regression Trees, henceforth Boosting. We simulated a consisting of 20 raters, 25 items, 30 patients, and thus in total 15,000 observations. This data set was split into a training set (90%) and a test set (10%). The performance of the three methods was evaluated by training the models on the training set and using the trained model to predict the outcomes for a test set. For the LTRM, we used the mode of the posterior predictive distribution as a point-prediction[2], whereas

[2] In this particular example, model predictions could also be interpreted as imputing a missing value. If these are regarded as missing observations rather than predictions, they should be modeled as unknown discrete parameters of the model (Ch. 8; Gelman et al., 2014). That way, uncertainty about these missing observations is propagated into the parameters. Although we did not sample the missing observations from the joint posterior distribution, the code in the online appendix does show how to do this.

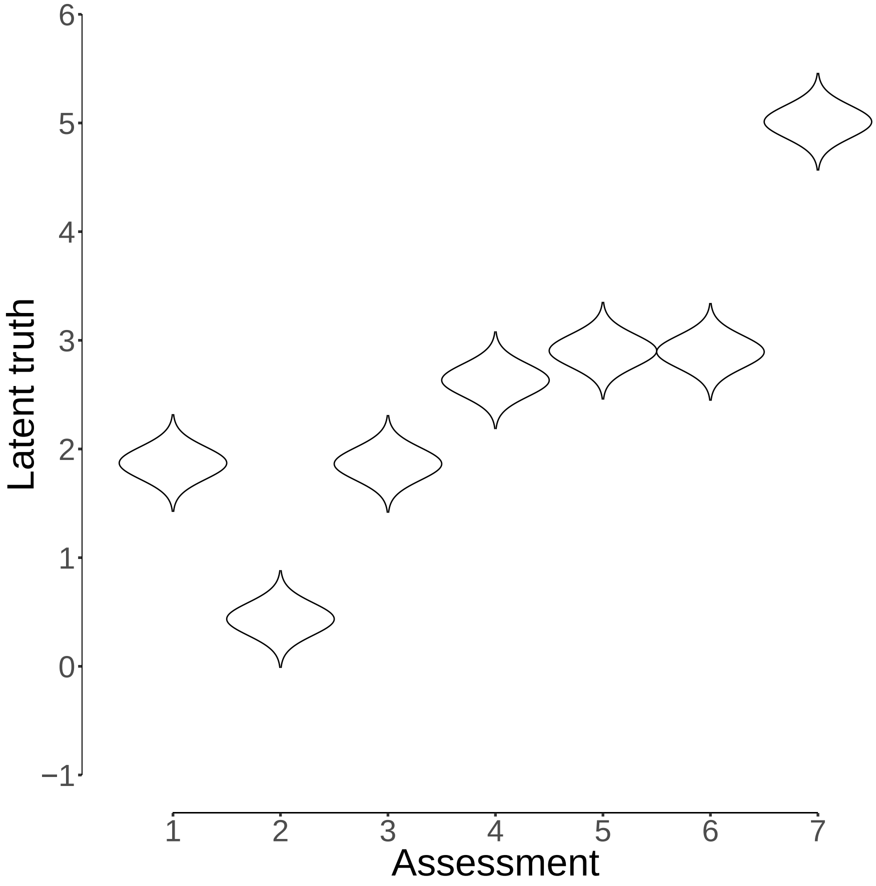


Figure 5: Example of how a patient's progress on a single construct could be monitored across measurement occasions.

for Random Forest and Boosting, we used the majority vote. Rather than the commonly used mean (which predictions lie outside of the ordinal scale), we used the observed mode of all observations for the same rater, item, or patient (i.e., mode$(x_{-r,ip}, x_{r,-i,p}, x_{ri,-p})$).

A technical difference between this approach and the previous simulation is that model predictions are effectively treated as missing data. In terms of implementation, a complete data set, where every patient is scored on each item by all raters, is represented as an array of dimensions $R \times I \times P$ (and possibly a dimension for measurement occasion). However, it is easier to handle an incomplete data set by representing the data as a matrix of $(R \times I \times P)$ rows and 4 columns (i.e., long format).[3] In this format, the first column indicates

[3] For a Stan implementation the long format is even required as missing values must be

the outcome, the second the rater, the third the item, the fourth the patient, and subsequent columns indicate rater and patient covariates.

We quantified prediction error by calculating the confusion matrix, a frequency table of predicted versus hold-out responses where the diagonal indicates the number of correct responses. Prediction accuracy is defined as the number of correct responses divided by the total number of responses.

Table 1: Prediction accuracy for the LTRM, Random Forest, Boosting, and the sample mode. The LTRM outperforms all other methods.

Method	Prediction Accuracy
LTRM	0.52
Random Forest	0.40
Boosting	0.35
Sample Mode	0.41

Although Random forest and Boosting are excellent black box algorithms for prediction, with default settings these algorithms initially performed worse than the observed mode. Therefore, we optimized the hyperparameters of Random Forest and Boosting (Kuhn et al., 2008).

Given that the data were generated by the LTRM, it comes as no surprise that it predicts more accurately than the other methods. However, even though data generated from the LTRM is likely a gross simplification of reality, the results show that black box machine learning methods perform inadequately as their predictive power does not exceed that of the sample mode. It is likely that the data at hand are ill-suited for black-box machine learning methods as they have difficulty capturing the hierarchical structure of the data which contains most of the information in this scenario. Instead, if a lot of background information about patients and raters is available, this could likely improve their performance.

Discussion

In this paper, we extended the Cultural Consensus model developed by Anders and Batchelder (2015) to be suited for data often encountered in psychiatric detention centers. The original model was suited for data from a single patient and we extended this to multiple patients, latent constructs, and patient and rater specific covariates. The benefit of this approach is that we can obtain estimates for e.g., a patients aggressiveness, while accounting for rater bias, item-specific measurement error, and a patient's criminal offense.

Although the LTRM provided better predictions than black-box machine learning approaches, this is likely because the data were simulated from the LTRM. It seems more reasonable that an optimal method for prediction would combine results from the LTRM with some machine learning approach. For

handled explicitly, unlike e.g., JAGS.

example, augmenting a Random forest model with features based on psychological theories resulted in improved predictions of human decisions (Plonsky et al., 2019; Plonsky, Erev, Hazan, & Tennenholtz, 2017). However, machine learning approaches, despite their predictive power, result in uninterpretable models which may not always be desirable in practice.

Recommendations for clinical practice

To successfully apply the LTRM model in practice, the data should meet several minimum requirements. This is to obtain some minimum quality

Evaluations should be recorded and stored long-term (e.g., multiple years). Ideally, ratings are obtained with high frequency. Differences between raters should be minimized, for instance by training staff. Additional information about patients, such as the reason of incarceration, should be added to the model.

Limitations

In the LTRM, we assumed that the factor structure is known. However, in practice, this need not be the case. Estimating this structure from the data is possible, however, such an endeavor may shift the focus of the LTRM to retrieving the latent factor structure, rather than interpreting the results of the model on an individual level.

To sum up, we introduced a model called the LTRM that is suitable for data typical in psychiatric detention centers. The model accounts for individual differences between raters, items, and patients. In a simulation, we have shown that the LTRM outperforms a the observed mode in terms of predictive power. Finally, we have provided recommendations for clinical practitioners who wish to apply the LTRM in practice.

References

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452–469.

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.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)*. Boca Raton (FL): Chapman & Hall/CRC.

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

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. (2019). gbm: Generalized boosted regression models. Retrieved from https://CRAN.R-project.org/package=gbm (R package version 2.1.5)

Kuhn, M., et al. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., ... others (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Thirty-first aaai conference on artificial intelligence.*

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.

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019, May 08). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*. Retrieved from https://doi.org/10.3758/s13428-019-01231-3 doi: 10.3758/s13428-019-01231-3

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. doi: 10.18637/jss.v077.i01

Parameter Recovery

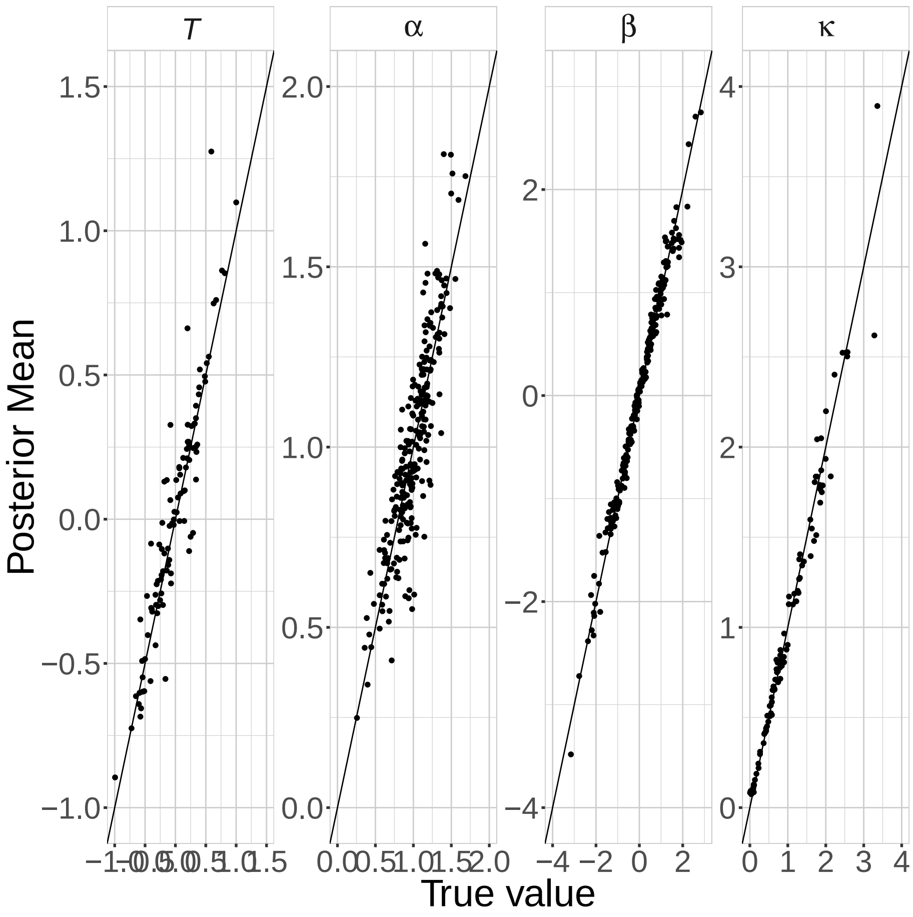


Figure 6: Parameter recovery for the model displayed in Figure 1