Augmenting Predictive Models in Forensic Psychiatry with Cultural Consensus Theory

Don van den Bergh[1], Erwin Schuringa[2], and Eric-Jan Wagenmakers[1]

[1]Department of Psychological Methods, University of Amsterdam
[2]Forensic Psychiatric Centre Dr. S. van Mesdag

Contents

Abstract

1 Introduction

The mental health and forensic risk factors of patients in forensic psychiatric hospitals is regularly monitored with methods such as Routine Outcome Monitoring (de Beurs et al., 2011).

Forensic psychiatric hospitals routinely assess the

Forensic psychiatric hospitals monitor the mental health and forensic risk factors of their patients at regular intervals, typically using a method such as Routine Outcome Monitoring (de Beurs et al., 2011).

2 Data of 105 patients in the Mesdag clinic

2.1 How it was collected

2.2 IFTE

The data were collected using a Routine Outcome Monitoring instrument called the the Instrument for Forensic Treatment Evaluation (IFTE). The IFTE consists of 22 items, of which 14 items are criminogenic need indicators of the Dutch risk assessment instrument HKT-R (Spreen et al., 2013), five items were designed in consultation with psychologists and psychiatrists, and three items are based on the Atascadero Skills Profile (Vess, 2001). The 22 items can be grouped into three factors, Protective behaviors, Problematic behaviors, and Resocialization Skills. All items are scored on a 17 point scale.

2.3 Descriptive Statistics

- Amount of patients, raters, and items.
- Sparsity of patient-rater combinations.

3 Cultural Consensus Theory

Cultural Consensus Theory (CCT) sometimes called also known as “test theory without an answer key” (Batchelder & Romney, 1988), is a method to discover the ”true answer“ for items from the consensus among the responses. For example, suppose a mentally ill patient is scored by multiple raters on aggressiveness. Multiple scores are obtained that need to be aggregated to arrive at a single score for this patient. The naive solution is to average these scores. However, as shown in Figure 1 averaging may lead to severely biased estimates.

DB: Figure here of score means versus true latent parameter for data simulated under the CRM.

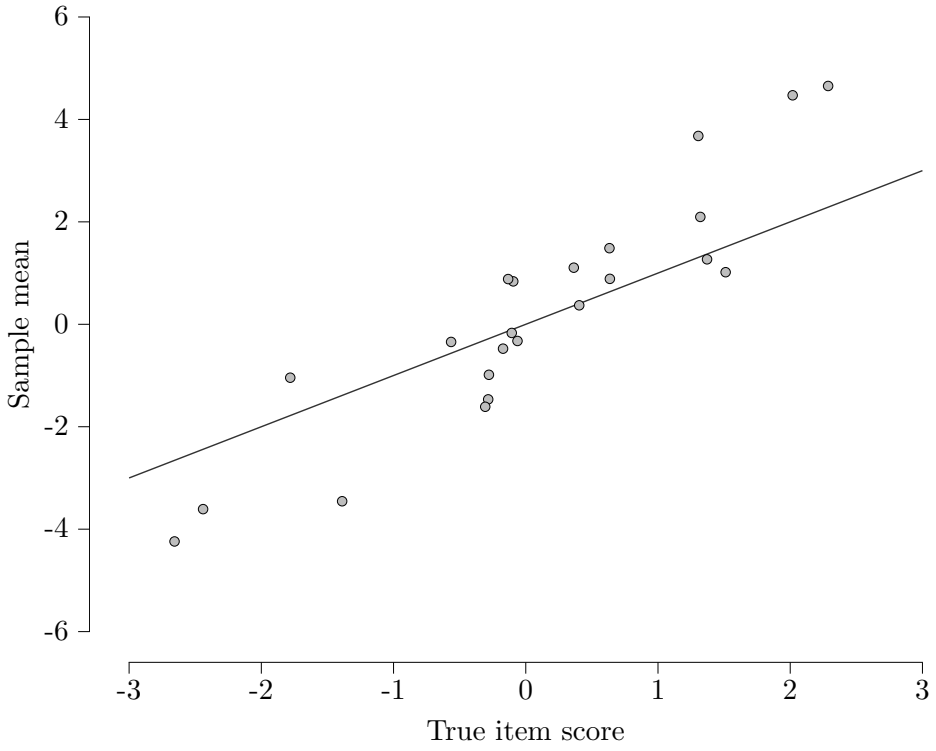


Figure 1: True item scores (x-axis) versus the sample mean across raters (y-axis).

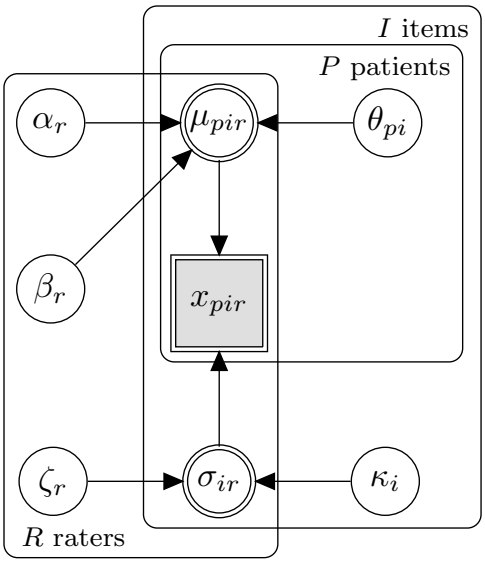
The average score disregards all additional information that is available. It ignores the individual differences between raters, for example, this assumes that all psychiatrists score aggressiveness in the same way, and group differences among raters, for example, there is no difference in scores by psychiatrists as opposed to clinicians, or other staff member. In addition, the average ignores any additional information about the patient at hand, such as the committed crimes and diagnoses.

Cultural consensus theory (CCT) provides a model-based framework for pooling information from multiple raters to form a consensus (Anders et al., 2014). There exist a variety of CCT models, each applicable to different types of data. For example, the General Condorcet model (Batchelder & Romney, 1986) applies to dichotomous data, the Latent Truth Rater model (Anders & Batchelder, 2015) is suited for ordinal data, and the Continous Response model (Anders et al., 2014).

small description

3.1 The Continuous Response Model

The Continuous Response Model (CRM) is a CCT model for continuous data (Anders et al., 2014). Figure 2 shows a graphical model of the CRM with a few adjustments.



$$
\begin{aligned}
x_{pir} &\sim \text{Normal}\left(\mu_{pir},\, \sigma^2_{pir}\right) \\
\mu_{pir} &= \alpha_r \theta_{pi} + \beta_r \\
\sigma_{ir} &= \zeta_r \kappa_i \\
\theta_{pi} &\sim \text{Normal}\left(\mu_\theta,\, \sigma^2_\theta\right) \\
\log \kappa_i &\sim \text{Normal}\left(\mu_\kappa,\, \sigma^2_\kappa\right) \\
\log \alpha_r &\sim \text{Normal}\left(\mu_\alpha,\, \sigma^2_\alpha\right) \\
\beta_r &\sim \text{Normal}\left(\mu_\beta,\, \sigma^2_\beta\right) \\
\log \zeta_r &\sim \text{Normal}\left(\mu_\zeta,\, \sigma^2_\zeta\right)
\end{aligned}
$$

Figure 2: Graphical model of the Continuous Response Model for multiple patients.

Here, x_{pir} is the score given to patient p on item i by rater r. This score is assumed to be normally distributed with mean μ_{pir} and standard deviation σ_{ir}. The mean is composed of a patients true score on a particular item θ_{pi} and a rater-specific scale α_r and shift β_r. The standard deviation is comprised of an item-specific difficulty κ_i and a rater-specific competence ζ_r. The ratio of the item difficulty and rater competence determines how precise an answer is retrieved.

There are three differences in Figure 2 compared to the CRM as described in Anders et al. (2014). First, we do not allow for multiple cultural truths among the raters. In our application the raters are professionally trained and we believe the remaining rater parameters suffice to capture differences between raters. Second, (Anders et al., 2014) considered two levels of nesting, items and respondents, whereas we consider three levels of nesting, patients, items, and raters. By dropping the patient indices one

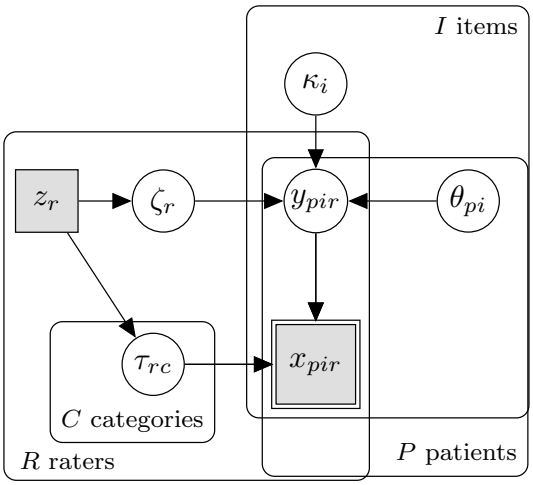
Third, we adjusted the prior distributions opposed

Different cultural truths allows the "true answer" on an item to vary across patients Another difference with respect to our previous work (van den Bergh et al., 2020) is that the item difficulty κ does not vary across patients. During simulations with a similar ratio of patients to raters, we found that this parameter cannot be reliably estimated. Therefore we opted

to constrain this parameter across patients.

3.2 The Latent Truth Rater Model

The Latent Truth Rater Model (LTM) is a CCT model for ordinal data (Anders & Batchelder, 2015). Previously, we extended the LTM to handle data from multiple patients (van den Bergh et al., 2020) and Figure 3 shows the LTM for multiple patients.



$$
\begin{aligned}
x_{pir} &= \begin{cases} 1 & \text{if } y_{pir} \leq \tau_{r1} \\ c & \text{if } \tau_{r,c-1} < y_{pir} \leq \tau_{rc} \\ C & \text{if } y_{pir} > \tau_{r,C-1} \end{cases} \\
y_{pir} &\sim \text{Logistic}\left(\theta_{pi},\ \kappa_i/\zeta_r\right) \\
\tau_{rc} &\sim \text{Normal}\left(0,\ 1\right) \\
\theta_{pi} &\sim \text{Normal}\left(\mu_\theta,\ \sigma^2_\theta\right) \\
\kappa_i &\sim \text{Gamma}\left(\mu^2_\kappa/\sigma^2_\kappa,\ \mu_\kappa/\sigma^2_\kappa\right) \\
\zeta_r &\sim \text{Gamma}\left(\mu^2_{\zeta_r}/\sigma^2_{\zeta_r},\ \mu_{\zeta_r}/\sigma^2_{\zeta_r}\right)
\end{aligned}
$$

Figure 3: Graphical model of the Latent Truth Rater Model for multiple patients. Note that the thresholds $\boldsymbol{\tau}_r$ are constrained to be ordered, that is, for all raters we have $\tau_{r1} \leq \cdots \leq \tau_{rc} \leq \cdots \leq \tau_{r,C-1}$.

Here, x_{pir} is the observed score given to patient p on item i by rater r. This score is assumed to be deterministically generated from a continuous latent appraisal y_{pir} that is discretized to an ordinal scale by the thresholds τ_{rc}. In particular, we have that

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

Since the appraisal score is latent, the deterministic function above implies the following probabilistic model over the observed scores:

$$
P(x_{pir} \mid y_{pir}, \boldsymbol{\tau}_r) = \begin{cases} 1 - F\left(y_{pir} - \tau_{r1}\right) & \text{if } x_{pir} = 1, \\ F\left(y_{pir} - \tau_{r,c-1}\right) - F\left(y_{pir} - \tau_{rc}\right) & \text{if } 1 < x_{pir} < C, \\ F\left(y_{pir} - \tau_{r,C-1}\right) & \text{if } x_{pir} = C. \end{cases}
$$

where $F\left(\right)$ is the logistic cumulative distribution function.[1]

[1]Note that this choice is arbitrary and that it is possibly to use any continuous cumulative distribution function.

Next, we explain how the latent appraisals and thresholds come about. The appraisals are draws from a logistic distribution with location θ_{pi}, the true score for patient p on item i. The scale of the logistic distribution is the ratio of the item difficulty κ_i to the rater competence ζ_r. A higher item difficulty means that the appraisals are more noisy, which leads to a more dispersed probability distribution over possible scores. Conversely, a higher rater competence means that the appraisals are less noisy, which leads to a more concentrated distribution over the outcomes. There are $C-1$ ordered thresholds for each rater, which are assigned a standard normal prior for identification purposes.

There are two differences in the model specification above compared to our previous work (van den Bergh et al., 2020). First, we previously modeled the thresholds using two rater specific parameters. However, in simulations we noticed that these two parameters provide too little flexibility when the ordinal scale consists of 18 categories, as in the data at hand. Therefore we decided to model the thresholds individually. This complicates interpreting the differences between the thresholds across raters, however, that is also not the goal of this paper. Second, we previously allowed the item difficulty parameter to vary across patients, which captures that some items may be more difficulty or easy to assess for some patients (e.g., some patients may cooperate more than others). This parameter is mainly informed by the number or raters and there needs to be a sufficient amount of raters that score each patient for a reliable estimate. However, we noticed when simulating data with a ratio of raters to patients like in the data at hand that there are simply too few observations to reliably estimate the deviations in item difficulty across patients.

DB: We could also look at a patient specific "cooperativeness" parameter. Rather than doing κ_{pi}, which introduces $P \times I$ parameters, the logistic scale would become $\frac{\kappa_i}{\zeta_r \times \text{Cooperativeness}_p}$. This introduces an additional P parameters.

3.3 Augmenting Logistic Regression with the CRM

In a next step, we use the information from the LTM to predict violent behavior. We use logistic regression to predict violent behavior. Figure 4 shows a graphical model of the logistic regression combined with the LTM. The latent truth for each patient on each item is seen as a covariate in the logistic regression model.

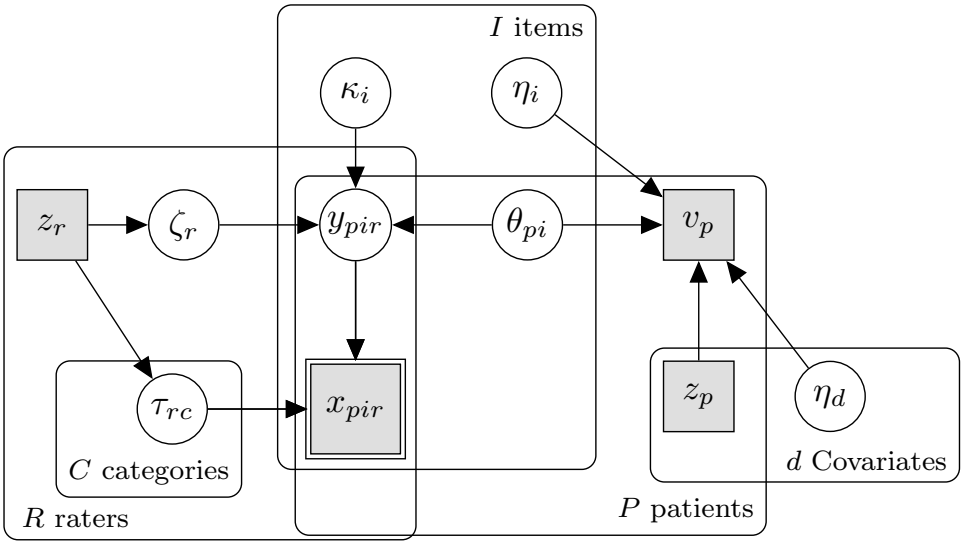


Figure 4: Graphical model of Logistic Regression Augmented with the Latent Truth Rater Model.

3.4 The Latent Truth Rater Model

3.5 Simulation Study

3.5.1 Do we actually want to report this?

3.6 Machine Learning Alternatives

This section discusses the four alternative considered, naive logistic regression, random forest, boosted regression trees.

3.6.1 Missing value imputation and data reduction

The four methods described above are designed for purely rectangular data. That is, each row of the data set contains one outcome (aggressive or not) and a number of predictors. However, the raw data contains missing valu This means that some preprocessing of the data is needed before these methods can be applied. Here we discuss missing data handling and the data reduction techniques applied.

Missing values The four alternative methods do not handle missing values. Although listwise deletion is an option, this results in rather We use the R package `mice` (van Buuren & Groothuis-Oudshoorn, 2011) to impute missing values. Rather than imputing one single value for each missing observation, `mice` imputes multiple, which leads to multiple data sets. While this increases the computational complexity of the procedures, it also propagates uncertainty about the missing observations to the results.

Data reduction After imputing missing values in the data, it remains problematic that there are multiple raters

4 Empirical Example

4.1 Data set about Forensic Psychiatric

doei

4.2 Discussion

Summary

Hybrid is probably most predictive.

4.3 Limitations

References

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

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In G. O. (B. Grofman (Ed.), *Information pooling and group decision making: Proceedings of the second University of California Irvine conference on political economy* (pp. 103–112). JAI Press Greenwich, CN.

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

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J. A., Giltay, E. J., van Noorden, M. S., van der Lem, R., van Fenema, E., & Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*(1), 1–12.

Spreen, M., Brand, E., Ter Horst, P., & Bogaerts, S. (2013). Handleiding hkt-r. *Historische Klinische Toekomst–Revisie*.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://www.jstatsoft.org/v45/i03/

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry*, *12*(3), 592–609.