

Augmenting Predictive Models in Forensic Psychiatry with Cultural Consensus Theory

Don van den Bergh¹, Erwin Schuringa², and Eric-Jan Wagenmakers¹

¹Department of Psychological Methods, University of Amsterdam

²Forensic Psychiatric Centre Dr. S. van Mesdag

Contents

1	Introduction	1
2	Data of 105 patients in the Mesdag clinic	2
2.1	How it was collected	2
2.2	IFTE	2
2.3	Descriptive Statistics	2
3	Cultural Consensus Theory	2
3.1	The Latent Truth Rater Model	3
3.2	Augmenting Logistic Regression with the CRM	5
3.3	The Latent Truth Rater Model	5
3.4	Simulation Study	5
3.4.1	Do we actually want to report this?	5
3.5	Machine Learning Alternatives	5
3.5.1	Missing value imputation and data reduction	6
4	Empirical Example	6
4.1	Data set about Forensic Psychiatric	6
4.2	Discussion	7
4.3	Limitations	7

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 (Sprenen 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 LTM.

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,

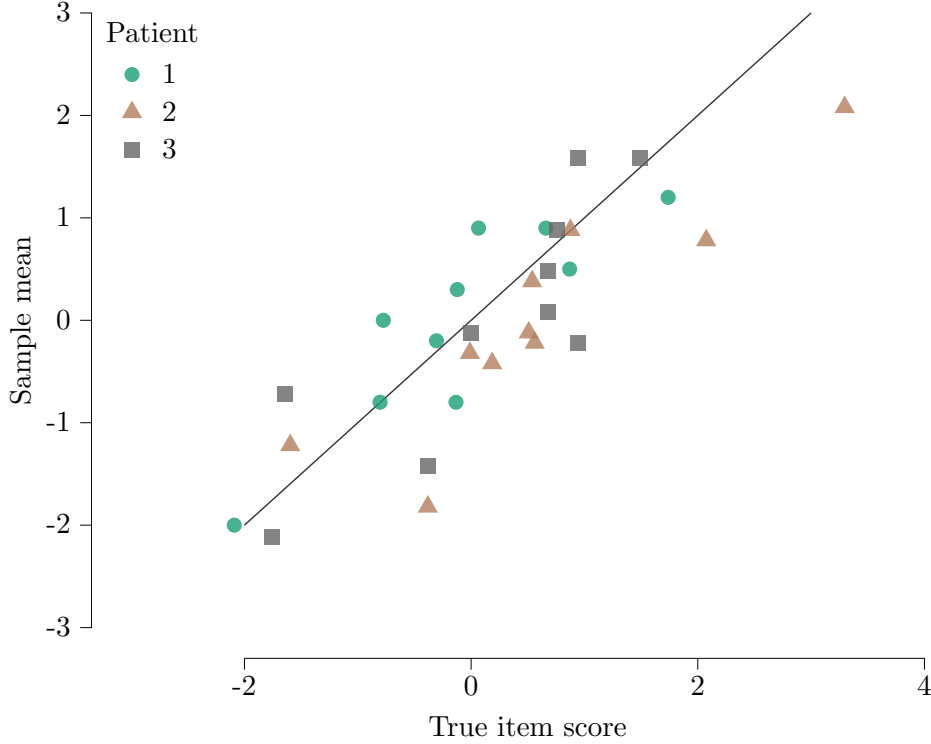


Figure 1: True item scores (x-axis) versus the sample mean across raters (y-axis).

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 Continuous Response model (Anders et al., 2014).

small description

3.1 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 2 shows the LTM for multiple patients.

Here, x_{pir} is the observed score given to patient p on item i by rater r .

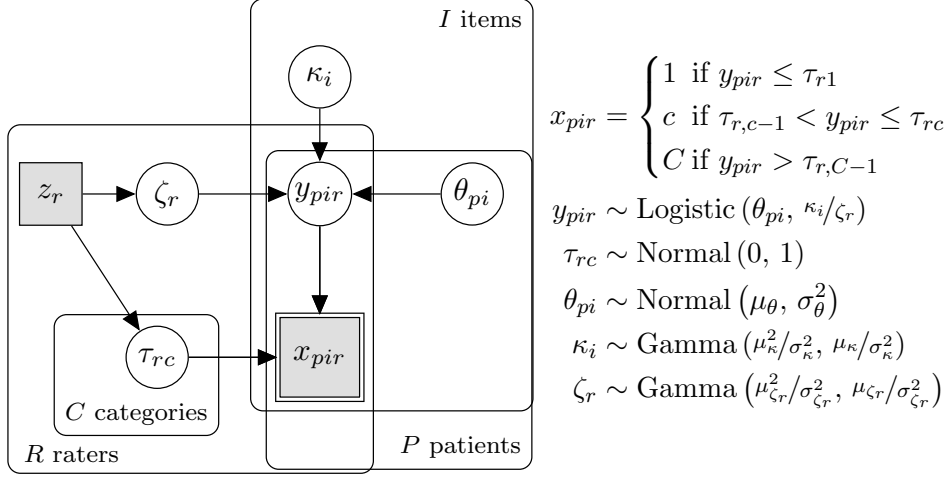


Figure 2: Graphical model of the Latent Truth Rater Model for multiple patients. Note that the thresholds τ_r are constrained to be ordered, that is, for all raters we have $\tau_{r1} \leq \dots \leq \tau_{rc} \leq \dots \leq \tau_{r,C-1}$.

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:

$$P(x_{pir} | y_{pir}, \tau_r) = \begin{cases} 1 - F(y_{pir} - \tau_{r1}) & \text{if } x_{pir} = 1, \\ F(y_{pir} - \tau_{r,c-1}) - F(y_{pir} - \tau_{rc}) & \text{if } 1 < x_{pir} < C, \\ F(y_{pir} - \tau_{r,C-1}) & \text{if } x_{pir} = C. \end{cases}$$

where $F()$ is the logistic cumulative distribution function.¹

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

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

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 Mesdag data. 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 of 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.

3.2 Augmenting Logistic Regression with the LTM

In a next step, we use logistic regression for to predict violent behavior, where we use the results from the LTM as additional predictors. We do so in a fully Bayesian approach, that is, we constructed a joint model for the violent behavior and then patient ratings.² Figure 3 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.

3.3 The Latent Truth Rater Model

3.4 Simulation Study

3.4.1 Do we actually want to report this?

3.5 Machine Learning Alternatives

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

²An alternative is to fit the LTM separately and then use e.g., the posterior means in a second step in a logistic regression. While this would be more computationally efficient, it would also ignore the uncertainty in the analysis of the patient ratings.

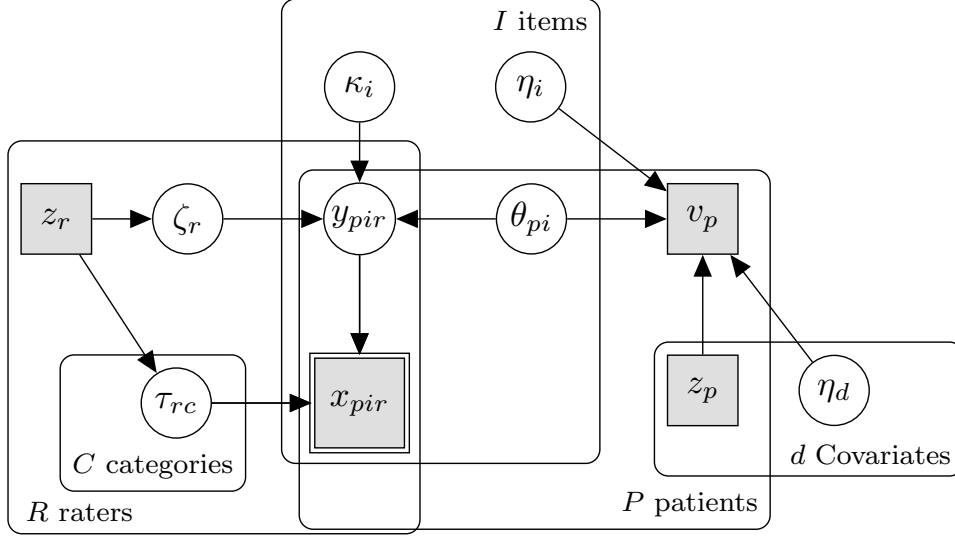


Figure 3: Graphical model of Logistic Regression Augmented with the Latent Truth Rater Model.

3.5.1 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 values. 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.