Bayesian probabilistic extensions of a deterministic classification model[1]

Iwin Leenen and Iven Van Mechelen
Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium

Andrew Gelman
Department of Statistics, Columbia University, NY 10027, New York

[1]The authors gratefully acknowledge Brian Junker, Herbert Hoijtink, and William Browne for helpful comments on an earlier draft of this paper, and Johannes Berkhof for helpful discussions.

This work was supported in part by the Research Fund of K.U.Leuven, Grant OT/96/10, and the U.S. National Science Foundation Grant SBR-9708424.

Summary

This paper extends deterministic models for Boolean regression within a Bayesian framework. For a given binary criterion variable Y and a set of k binary predictor variables $X_1, \ldots, X_k$, a Boolean regression model is a conjunctive (or disjunctive) logical combination consisting of a subset S of the X variables, which predicts Y. Formally, Boolean regression models include a specification of a k-dimensional binary indicator vector $(\theta_1, \ldots, \theta_k)$ with $\theta_j = 1$ iff $X_j \in S$. In a probabilistic extension, a parameter π is added which represents the probability of the predicted value $\hat{y}_i$ and the observed value y_i differing (for any observation i). Within a Bayesian framework, a posterior distribution of the parameters $(\theta_1, \ldots, \theta_k, \pi)$ is looked for. The advantages of such a Bayesian approach include a proper account of the uncertainty in the model estimates and various possibilities for model checking (using posterior predictive checks). We illustrate this method with an example using real data.

Keywords: Bayesian Estimation, Boolean Regression, Logical Rule Analysis, Posterior Predictive Checks

1 Introduction

In many research areas, prediction problems are considered with the predictors and/or criteria being binary variables. As a result, a number of models and associated techniques have been developed to examine the relations in this type of data, including instantiations of the generalized linear model. For example, in a logistic regression model with binary variables, the logit of the probability that the criterion variable assumes either of the two possible values is a linear function of a number of predictors. In many relevant cases, though, one aims at finding the sufficient and/or necessary conditions for a criterion to occur; hence, this makes the generalized linear modeling approach, which assumes a compensatory association rule, less appropriate from a theoretical point of view. In medical diagnoses, for example, assigning a disease to a given patient is often based on considering a list of necessary and sufficient conditions; as another example, so-called classical theories on categories and concepts assume that assignment to a category is based on the presence of a set of singly necessary and jointly sufficient attributes.

In search of necessary (or sufficient conditions), a Boolean regression model (Van Mechelen, 1988; Van Mechelen & De Boeck, 1990) may be helpful as it identifies for a given binary criterion and a given set of binary predictors a subset of the predictors that are conjunctively (resp. disjunctively) combined to predict the value on the criterion variable. Besides applications in the social sciences (McKenzie, Clarke, & Low, 1992; Ragin, Mayer, & Drass, 1984; Van Mechelen & De Boeck, 1990), techniques related to Boolean regression have been studied in discrete mathematics and in the context of the

design of switching circuits in electronics (Biswas, 1975; Halder, 1978; McCluskey, 1965; Sen, 1983). In these applications, more complex rules, such as disjunctive combinations of conjunctions (or vice versa), are also considered.

Boolean regression has initially been formulated as a deterministic model. Existing algorithms for Boolean regression aim at finding a subset of the predictors which minimizes the number of prediction errors (Van Mechelen, 1988). However, at least three drawbacks can be met with the approach of finding a single best solution. First, in many empirical applications, several different subsets of the predictors may fit the data (almost) equally well, whereas from a substantive viewpoint they may be quite different. Second, it is not obvious how to draw statistical inferences about the size of the prediction error as the prediction error associated with the single best solution probably underestimates the true model error (because the algorithm aims at minimizing the number of prediction errors). Third, the deterministic model does not provide any tools for model checking due to the fact that in the model the relation to the data is not specified. Hence, a method which gives insight in several concurring models and in the level of uncertainty associated with them, is of great interest.

Therefore, the present paper extends the model of Boolean regression within a Bayesian framework. Bayesian statistics can be considered a natural conceptual framework for exploring the likelihood of several possible concurring models for a given data set. The model extension presented here follows the general recipe proposed by Gelman, Leenen, Van Mechelen, and De Boeck (1999), which brings most of the tools that are available for stochastic models within the realm of deterministic models (like the model of Boolean regression).

The remainder of the paper is organized as follows: Section 2 recapitulates the deterministic model of Boolean regression. In Section 3, the stochastic extension is presented and estimation and checking of the model within a Bayesian framework is discussed. In Section 4 an example on definitions of human emotions illustrates the application of the new model to real data. Section 5 deals with possible extensions and contains some concluding remarks.

2 The Deterministic Boolean Regression Model

2.1 Model formulation

Consider an $n \times k$ binary matrix X, which denotes the observations for n units on k explanatory variables $X_1, \ldots, X_j, \ldots, X_k$, and a binary vector $y = (y_1, \ldots, y_i, \ldots, y_n)$, which contains the observed values for the n units of a criterion variable Y. Boolean regression, then, specifies a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$ with $\theta_j \in \{0, 1\}$ $(j = 1, \ldots, k)$ which is subsequently combined with X to give a binary vector $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$ of predicted values

of the criterion. Both a disjunctive and a conjunctive variant of the model exist, which differ in the way that θ and X are combined to get $\hat{y}$. In a conjunctive model,

$$\hat{y}_i \equiv \hat{y}(\theta, X)_i = \prod_{j|\theta_j=1} x_{ij}, \tag{1}$$

whereas in the disjunctive variant:

$$\hat{y}_i \equiv \hat{y}(\theta, X)_i = 1 - \prod_{j|\theta_j=1} (1 - x_{ij}) \,. \tag{2}$$

Despite their substantive difference, conjunctive and disjunctive models are dual models: A comparison of Eq. (1) and Eq. (2) shows that if a conjunctive model fits some data set X and y then simultaneously the disjunctive model fits the complemented data X^C and y^C, and vice versa, where $x_{ij}^C = 1 - x_{ij}$ and $y_i^C = 1 - y_i$ $(i = 1, \ldots, n; j = 1, \ldots, k)$. As a result, only one of the variants needs to be considered; in this paper, we focus on the conjunctive model and, unless otherwise stated, any $\hat{y}_i$ is calculated as in Eq. (1).

Boolean regression being a deterministic model does not include a specification of the relation between the observed y and the predicted $\hat{y}$; strictly speaking, the model requires them to be equal. Hence, whenever an observation i exists for which y_i and $\hat{y}_i$ are discrepant (i.e., $y_i \neq \hat{y}_i$), the model should be rejected. In practical applications of the model, though, one allows for prediction errors, and algorithms for fitting the model aim at finding θ with the minimal number of discrepancies:

$$D(y, \theta) = \sum_{i=1}^{n} \left[y_i - \hat{y}(\theta, X)_i \right]^2 . \tag{3}$$

As y_i and $\hat{y}_i$ are binary, (3) can be considered both a least squares and a least absolute deviations (or, more generally, a least L_p-norm, $p > 0$) loss function.

2.2 Model estimation

To find a θ that minimizes $D(y, \theta)$, two strategies have been proposed. Most algorithms (Mickey, Mundle, & Engelman, 1983; Van Mechelen, 1988; Van Mechelen & De Boeck, 1990) use a greedy heuristic which initializes the entries in θ to 1 and successively changes the value of some entry θ_j into 0, each time selecting that θ_j for which the change yields the largest decrease in number of discrepancies, until changing any of the remaining θ_j's does not further improve the solution.

Recently, Leenen and Van Mechelen (1998) have proposed a branch-and-bound algorithm that guarantees that a solution with minimal value on $D(y, \theta)$ is found. This algorithm passes through a tree, making extensively use of the property that in a conjunctive model changing an arbitrary entry

θ_j from 1 into 0 does not decrease the number of false negatives (a false negative being defined as an observation i for which the predicted value $\hat{y}_i$ equals 0 and the observed value y_i equals 1). In many cases, the latter property allows the algorithm to apply branching and bounding to a large extent, thereby strongly reducing the processing time compared to an enumerative search among all possible solutions.

2.3 Model checking

The goodness of fit of the deterministic model can be summarized into a number of descriptive statistics, including proportion of discrepancies, Jaccard's goodness-of-fit statistic (Sneath & Sokal, 1973; Tversky, 1977), and Van Mechelen and De Boeck's (1990) $\hat{\lambda}_p$, which indicates the amount of predictive gain by knowing the model over a prediction based on the marginal criterion probability only. However, these statistics are limited in that they measure the total goodness of fit and do not allow for a more in depth examination of the structure of the errors. Also, only rules of thumb are available to decide on whether or not a solution is "sufficiently good".

3 Bayesian Boolean Regression

3.1 Model formulation

Allowing for discrepancies reveals the implicit assumption of a stochastic model underlying the deterministic model. A natural extension of the model may therefore be considered that explicitly includes the possibility of a prediction error.

The stochastic extension implies the addition of a Bernoulli-like process to the deterministic model, which accounts for the values on the criterion variable possibly changing from 0 into 1 or vice versa. For this purpose, a new parameter π is added to the model, which is the expected error rate of the model and which is assumed to be identical across observations. Hence, for any observation i, it holds that:

$$\Pr(y_i = \hat{y}_i|\theta, \pi) = 1 - \pi. \tag{4}$$

(In the latter and all following equations, the dependence on X is not explicitly indicated because the predictor values are considered fixed.) Under local stochastic independence, it further holds that the likelihood of y under this model is:

$$p(y|\theta, \pi) = \pi^{D_\theta}(1 - \pi)^{n - D_\theta}.$$

For convenience, $D(y, \theta)$ is abbreviated to D_θ in formulas.

In a next step, the stochastic model is considered within a Bayesian framework, which provides tools for exploring the posterior distribution:

$$p(\theta, \pi|y) = \frac{p(y|\theta, \pi)\ p(\theta, \pi)}{p(y)}. \tag{5}$$

We will assume that θ and π have independent and uniform prior distributions. Uniform prior distributions imply a minimal extension of the already-existing deterministic model: In this case maximizing the likelihood (which implies minimizing the number of discrepancies) corresponds to finding the mode of the posterior distribution (Gelman et al., 1999). Alternatively, an informative prior distribution may be considered, particularly if some substantive knowledge on the parameters is available.
As shown in the Appendix, working out the posterior density yields:

$$p(\theta, \pi|y) = \frac{(n+1)\ \pi^{D_\theta}\ (1-\pi)^{n-D_\theta}}{\sum\limits_{\vartheta \in \Theta} \frac{1}{\binom{n}{D_\vartheta}}}, \tag{6}$$

where the sum in the denominator is over all 2^k values in the parameter space Θ. Clearly, evaluating this sum is feasible for small k only.

Often, one will be interested in the marginal posterior distribution of the θ parameter. In the Appendix it is shown that integrating out π in Eq. (6) results in:

$$p(\theta|y) = \frac{\frac{1}{\binom{n}{D_\theta}}}{\sum\limits_{\vartheta \in \Theta} \frac{1}{\binom{n}{D_\vartheta}}}. \tag{7}$$

The latter implies that two θ parameters that yield $\hat{y}$'s that are equally iscrepant with y have equal posterior probabilities. Furthermore, it follows that if θ has one discrepancy fewer than θ^* then the ratio of their marginal posterior probabilities equals:

$$\frac{p(\theta|y)}{p(\theta^*|y)} = \frac{n - D_\theta}{D_{\theta^*}}. \tag{8}$$

3.2 Model estimation

In this section we show how one can gain insight into the posterior distribution by drawing simulations with a Gibbs-Metropolis algorithm:

Step 0 As an initialization step, m estimates $\theta^{(s,0)}$ and m estimates $\pi^{(s,0)}$, $(s = 1, \ldots, m)$, are constructed as follows: $\theta^{(s,0)}$ is a random binary vector with $\Pr(\theta_j^{(s,0)} = 1) = 0.5$ $(j = 1, \ldots, k)$ and $\pi^{(s,0)}$ is given the value:

$$\pi^{(s,0)} \leftarrow \frac{D(y, \theta^{(s,0)}) + 1}{n + 2}.$$

We add 1 in the numerator and 2 in the denominator to avoid initial estimates of π being 0 or 1 (Gelman et al., 1999).

Step 1 We run m parallel sequences of a Metropolis algorithm, with $(\theta^{(s,0)}, \pi^{(s,0)})$ as the starting point for sequence s $(s = 1, \ldots, m)$. At each iteration t $(t = 1, 2, \ldots)$, the following substeps are executed for each sequence s:

1. A candidate value θ^* is constructed based on the value $\theta^{(s,t-1)}$ in the previous iteration. Therefore, first an integer $w^{(s,t)}$ from a discrete density (e.g., Poisson or binomial) is drawn with the restriction $1 \leq w^{(s,t)} \leq k$. Next, $w^{(s,t)}$ entries in $\theta^{(s,t-1)}$ are randomly selected and subsequently changed (from either 0 into 1 or 1 into 0) to obtain θ^*.
   This procedure for constructing θ^* technically corresponds to drawing from the following jumping distribution:
   $$J(\theta^*|\theta^{(s,t-1)}) = \frac{1}{\binom{k}{w}} p(w), \tag{9}$$
   where $w \equiv w^{(s,t)}$ represents the number of entries in θ^* that are changed from $\theta^{(s,t-1)}$ and $p(\cdot)$ is the (truncated) discrete density mentioned above. The jumping distribution returns the probability of considering the candidate θ^*, given the value of $\theta^{(s,t-1)}$ of the previous iteration.
   Clearly, J is symmetric: $J(\theta^*|\theta) = J(\theta|\theta^*)$ such that the resulting algorithm is of the Metropolis type.

2. The ratio of the posterior densities, or equivalently, the ratio of the likelihoods, is calculated:
   $$r = \frac{p(y|\theta^*, \pi^{(s,t-1)})}{p(y|\theta^{(s,t-1)}, \pi^{(s,t-1)})} = \left(\frac{1-\pi^{(s,t-1)}}{\pi^{(s,t-1)}}\right)^{D\left(y,\theta^{(s,t-1)}\right)-D(y,\theta^*)}.$$

3. Values are assigned to $\theta^{(s,t)}$ and $\pi^{(s,t)}$:
   $$\theta^{(s,t)} \leftarrow \begin{cases} \theta^* & \text{with probability } \min(1, r) \\ \theta^{(s,t-1)} & \text{otherwise} \end{cases}$$
   The value for $\pi^{(s,t)}$ is obtained by a draw from a Beta$(D(y, \theta^{(s,t)}) + 1, n - D(y, \theta^{(s,t)}) + 1)$ distribution.

These steps are repeated until the m sequences appear mixed. Gelman and Rubin's (1992) $\sqrt{\hat{R}}$ statistic may be used as a diagnostic instrument in monitoring the convergence.

Step 2 In order to obtain L posterior simulation draws, the procedure described in step 1 continues, after convergence of the sequences, for another L/m iterations. The latter draws in the m sequences are collected and will eventually constitute the set of simulation draws $\{(\theta^{(l)}, \pi^{(l)}) \mid (l = 1, \ldots, L)\}$ from the posterior distribution.

3.3 Model checking

A natural way for model checking in Bayesian statistics is using posterior predictive checks (see Berkhof, Van Mechelen, & Hoijtink, this issue; Gelman, Carlin, Stern, & Rubin, 1995; Gelman, Meng, & Stern, 1996; Rubin, 1984). A posterior predictive check implies a comparison between the observed data and data replicated under the model, on some pre-defined test variable. If the test value for the observed data systematically deviates from the test value for the replicated data, then the observed data are unlikely to originate from the model under study. Formally, we proceed with the next steps:

Step 3 For each of the L simulated posterior draws a replicated data set $y^{(l)}$ is simulated as follows: First, $\hat{y}^{(l)} = \hat{y}(\theta^{(l)}, X)$ is computed using Eq. (1), and, subsequently, the n components of $y^{(l)}$ are independently simulated from $\hat{y}^{(l)}$ based on Eq. (4) (with $\pi^{(l)}$ substituted for π).

Step 4 A test variable $T(y, \theta)$ is defined which summarizes some aspect of interest of the data or some discrepancy between model and data.

Step 5 The realized value $T(y, \theta^{(l)})$ for the observed data and the replicated value $T(y^{(l)}, \theta^{(l)})$ for the replicated data are computed for each of the L simulation draws.

Step 6 The realized value and the replicated value are compared to estimate the posterior predictive p-value as the proportion of the L simulations for which $T(y^{(l)}, \theta^{(l)}) > T(y, \theta^{(l)})$.

The model checking procedure presented here will be illustrated in the example.

4 Illustrative Application

4.1 Problem and Data

In this section we illustrate the new approach with an example in the field of defining emotion concepts. According to Wierzbicka (1992, p. 541), emotion concepts can be defined by a set of singly necessary and jointly sufficient *semantic primitives*, which are "terms of words which are intuitively understandable (nontechnical), and which themselves are not names of specific emotions or emotional states." Table 1 lists some of the semantic primitives she proposed. As her definitions of emotions are conjunctive combinations of semantic primitives, a Boolean regression model may be expected to appropriately describe the relation between semantic primitives (as predictors) and an emotion concept (as the criterion). Whereas Wierzbicka deals with explicit definitions (i.e., by experts), the present study considers implicit theories by laymen and evaluates whether these implicit theories are conjunctive combinations of semantic primitives as well.

Table 1: List of the (noncomplemented) predictors for the Boolean regression analyses in the application

Predictor	Semantic primitive
X_1	A person did something bad
X_2	I don't want this
X_3	I would want to change this
X_4	I would want to do something bad to somebody
X_5	I feel bad
X_6	Something bad happened
X_7	I would want that something didn't happen
X_8	I can't change the situation
X_9	Something good happened
X_{10}	I want something like this
X_{11}	I feel good
X_{12}	Somebody did something good
X_{13}	I don't want to change this
X_{14}	I would want to do something good for somebody

Five first-year psychology students of the University of Leuven were each asked to generate twenty different situations in which they had recently been involved and felt either angry, sad, grateful, or happy. Next, the subjects were asked to specify for each of the twenty situations they generated: (1) whether or not each of the 14 semantic primitives in Table 1 was true and (2) whether or not they experienced each of the 4 forementioned emotions: anger, sadness, gratitude, and happiness. In the analyses, the 5×20 situations were concatenated, resulting into $n = 100$ observations, and both the original and the complemented semantic primitives were included as predictors, eventually resulting in 28 predictors $(X_1, \ldots, X_{14}, \neg X_1, \ldots, \neg X_{14})$ and 4 criteria Y_{angry}, Y_{sad}, Y_{grateful}, and Y_{happy}. Because the results for both negative emotions, anger and sadness, were very similar, as were the results for both positive emotions, gratitude and happiness, only analyses with anger and happiness are presented in the following sections.

4.2 Deterministic analysis

Optimal conjunctive logical rules (i.e., with minimal number of discrepancies) were found using the previously discussed branch-and-bound algorithm (Leenen & Van Mechelen, 1998). For Y_{angry}, the best logical rule combines the complements of the predictors 9, 10, and 14: A person reports (s)he experiences anger in a given situation iff "it is not the case that something good happened *and* (s)he does not want something like this *and* (s)he does not

want to do anything good for somebody". Y_{happy} on the other hand is best predicted by the single predictor 9: A person reports (s)he feels happy iff "something good happened". Table 2 presents some goodness-of-fit indices for both optimal rules.

Table 2: Optimal logical rules for Y_{angry} and Y_{happy} and associated goodness-of-fit statistics as found by a deterministic analysis

Emotion	Optimal rule	% discrepancies	Jaccard index	$\hat{\lambda}_p$
Anger	$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{14}$	9	.80	.75
Happiness	X_9	6	.89	.88

4.3 Bayesian analysis

4.3.1 Model estimation

The procedure discussed in Section 3.2 was used to simulate the posterior distribution of (θ, π). For each criterion, we ran $m = 5$ sequences of the described Gibbs-Metropolis algorithm. In the jumping distribution (9) we used a (truncated) Poisson density with parameter 2 for $p(\cdot)$. After convergence, that is when Gelman and Rubin's (1992) $\hat{R}$-statistic was smaller than 1.1 for each of the parameters θ_j $(j = 1, \ldots, 28)$ and π, another 2000 runs in each sequence were executed, ending up with $L = 10000$ posterior draws for each criterion.

Table 3 gives the marginal distribution of the θ-parameters (or, equivalently, the conjunctive combinations) for anger and happiness, respectively. The results of the Bayesian analysis show that the rule found by the deterministic branch-and-bound algorithm may be one of several "best" solutions: For anger, (at least) five different conjunctive combinations have a minimal number of discrepancies and two have only 1 discrepancy more; for happiness, four conjunctive combinations do equally well and 12 logical rules have 1 discrepancy more. In our example, $n = 100$, so the % discrepancies in Table 3 is equal to the number of discrepancies. In the table, models with the same number of discrepancies have different computed posterior probabilities; this is entirely due to simulation variability. The wide range of available models that fit about equally well indicates that the stochastic extension can add a considerable amount of information to the deterministic analysis, as otherwise only a single rule might be considered. For this particular case, one may remark that most of the other rules merely add one or more predictors to the rule found by the deterministic algorithm. Even then, however, the Bayesian analysis gives more insight into the error rate associated with the best solutions and into which other values for θ can reasonably be considered

Table 3: Simulated posterior distribution of θ for Y_{angry} and Y_{happy}

Logical rule	Post. prob.	% discrep.
Angry		
$\neg X_{10} \wedge \neg X_{11} \wedge \neg X_{12} \wedge \neg X_{14}$	.189	9
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{14}$	.183	9
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{11} \wedge \neg X_{12} \wedge \neg X_{14}$	.183	9
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{11} \wedge \neg X_{14}$	.166	9
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{12} \wedge \neg X_{14}$	.164	9
$\neg X_{10} \wedge \neg X_{12} \wedge \neg X_{14}$	.019	10
$\neg X_{10} \wedge \neg X_{11} \wedge \neg X_{14}$	.015	10
other	$< .006$	≥ 11
Happy		
X_9	.204	6
$\neg X_4 \wedge X_9$	.185	6
$\neg X_1 \wedge \neg X_4 \wedge X_9$	.170	6
$\neg X_1 \wedge X_9$	.170	6
$\neg X_4 \wedge X_{11}$	.030	7
$\neg X_5 \wedge X_9$	.023	7
$\neg X_1 \wedge \neg X_5 \wedge X_9$	.017	7
$\neg X_1 \wedge \neg X_4 \wedge X_{11}$	.017	7
$\neg X_4 \wedge \neg X_5 \wedge X_9$	.017	7
$\neg X_4 \wedge \neg X_5 \wedge X_{11}$	.017	7
X_{11}	.016	7
$\neg X_1 \wedge \neg X_4 \wedge \neg X_5 \wedge X_{11}$	.016	7
$\neg X_1 \wedge \neg X_5 \wedge X_{11}$	.016	7
$\neg X_5 \wedge X_{11}$	.015	7
$\neg X_1 \wedge \neg X_4 \wedge \neg X_5 \wedge X_9$	.014	7
$\neg X_1 \wedge X_{11}$	.010	7
other	$< .004$	≥ 8

for the given data set. With respect to the error rate, it was found for anger that $\pi = .100$ and for happiness that $\pi = .072$.

For purposes of illustration, it would be interesting to compare the simulated posterior distribution in Table 3 with the exact posterior distribution. Although an exhaustive evaluation of all conjunctions is not impossible in this case[1], for convenience we will present a reanalysis of the criteria anger and happiness using only a subset of predictors: For the criterion anger, only five predictors showed up in the conjunctive rules with posterior probability over .10, namely: $\neg X_9$, $\neg X_{10}$, $\neg X_{11}$, $\neg X_{12}$, and $\neg X_{14}$. Similarly, for happi-

[1] In total, $3^{14} = 4,782,969$ combinations are to be considered: each of the 14 predictors is either positively present, negatively present, or not present.

ness, the logical rules with highest posterior mass only use a subset of the six predictors $\neg X_1$, $\neg X_4$, $\neg X_5$, X_9, X_{10}, and X_{11}. We will discuss the results of a reanalysis of both criteria with only the semantic primitives that appeared relevant as predictors. This allows us to conveniently compute the exact posterior distributions and to compare this exact distribution with the simulated distribution. As in the previous analysis, we use $m = 5$ sequences and, after convergence, we collect $L = 10000$ posterior draws for both criteria.

Table 4: Exact and simulated posterior distribution of θ for Y_{angry} and Y_{happy} using relevant predictors only

Logical rule	Exact post. prob.	Simulated post. prob.
Angry		
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{11} \wedge \neg X_{12} \wedge \neg X_{14}$	.189	.212
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{11} \wedge \neg X_{14}$	.189	.189
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{14}$	.189	.189
$\neg X_{10} \wedge \neg X_{11} \wedge \neg X_{12} \wedge \neg X_{14}$	.189	.178
$\neg X_9 \wedge \neg X_{10} \wedge \neg X_{12} \wedge \neg X_{14}$	.189	.170
$\neg X_{10} \wedge \neg X_{11} \wedge \neg X_{14}$	.021	.028
$\neg X_{10} \wedge \neg X_{12} \wedge \neg X_{14}$	.021	.017
other	$< .005$	$< .003$
Happy		
X_9	.200	.212
$\neg X_1 \wedge X_9$	.200	.195
$\neg X_4 \wedge X_9$	.200	.192
$\neg X_1 \wedge \neg X_4 \wedge X_9$	.200	.190
$\neg X_1 \wedge \neg X_5 \wedge X_{11}$	.015	.019
$\neg X_4 \wedge \neg X_5 \wedge X_9$	.015	.019
$\neg X_5 \wedge X_9$	.015	.018
$\neg X_1 \wedge X_{11}$	.015	.018
$\neg X_5 \wedge X_{11}$	.015	.017
$\neg X_1 \wedge \neg X_4 \wedge \neg X_5 \wedge X_{11}$	.015	.017
$\neg X_1 \wedge \neg X_4 \wedge X_{11}$	.015	.016
$\neg X_1 \wedge \neg X_5 \wedge X_9$	.015	.015
$\neg X_4 \wedge \neg X_5 \wedge X_{11}$	.015	.015
$\neg X_1 \wedge \neg X_4 \wedge \neg X_5 \wedge X_9$	.015	.013
X_{11}	.015	.013
$\neg X_4 \wedge X_{11}$	.015	.012
other	$< .003$	$< .002$

Table 4 displays the marginal posterior distributions of θ (exact and sim-

ulated) for anger and happiness respectively. The results show that the simulated distribution is always close to the exact distribution, from which we may conclude that the estimation procedure works fine.

4.3.2 Model checking

In this section, the posterior-predictive-check approach is illustrated for checking a particular assumption that implicitly underlies the model applied in the previous subsection. We assumed that π was constant across observations (see, Eq. (4)); hence, in the study discussed above, no differences among the five subjects involved are allowed. Or, stated otherwise, the subjects are assumed to apply the respective logical rules with equal accuracy. (Clearly, if the model is found to be wrong at this point, an extension may be considered giving each person h a private expected error rate π_h.)

Individual differences in error rate may be quantified by the variance in number of prediction errors between the five subjects. Therefore, a test variable $T(y, \theta)$ is defined as:

$$T(y,\theta) = \frac{\sum_{h=1}^{5} \left[\frac{D_h(y,\theta)}{20} - \frac{D(y,\theta)}{100} \right]^2}{5-1},$$

where D_h is the number of discrepancies between the 20-element y and $\hat{y}$ vectors of subject h. The larger the variation among subjects, the larger the value of T.

For both criteria anger and happiness, 10000 replicated data sets $y^{(1)}, \ldots, y^{(10000)}$ were simulated as described in Section 3.3 and both the realized value $T(y, \theta^{(l)})$ and the replicated value $T(y^{(l)}, \theta^{(l)})$ $(l = 1, \ldots, 10000)$ were calculated. Next, posterior predictive p-values were computed as the proportion of the 10000 simulations for which $T(y, \theta^{(l)}) > T(y^{(l)}, \theta^{(l)})$. For anger, the posterior predictive p-value equals .566 and for happiness, it equals .624, which is visualized in Figure 1, which plots the observed versus the replicated values on the test variable: Roughly half the number of points lie above and half the number of points below the first bisector. As a result, it is concluded that the posterior predictive check provides no evidence for individual differences in accuracy.

5 Conclusions and further work

In some cases, one may expect the probability of a false positive to differ from the probability of a false negative prediction error. For example, in a psychiatric context, in a study on the implicit decision rule of clinicians with respect to the assignment of inpatients to a special suicide prevention program (psychiatric symptoms being the predictors), false positives may be more likely to occur than false negatives for reasons of caution. As discussed by Gelman

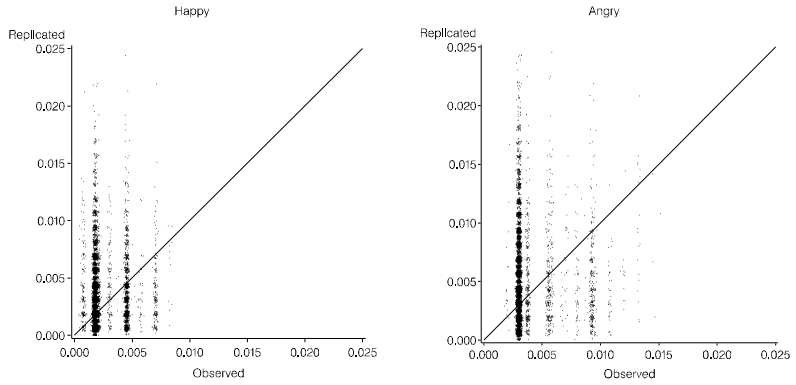

Figure 1: Plot of the realized $T(y, \theta^{(l)})$ versus the replicated $T(y^{(l)}, \theta^{(l)})$. The x and y coordinates are jittered by adding normal random numbers to each point's coordinates (with standard deviation .001) in order to display multiple values.

et al. (1999), the probabilistic model proposed here can be straightforwardly expanded by allowing different error rates π_0 and π_1 for responses predicted to by 0 and 1, respectively.

In contrast with most other Bayesian generalizations of deterministic disjunctive/conjunctive models (Gelman et al., 1999), for the Boolean regression model there is no need to restrict π (or π_0 and π_1) to be smaller than 0.5. Moreover, allowing π to cover the complete range from 0 to 1 may be helpful in distinguishing between disjunctive and conjunctive association rules. Indeed, from our discussion on the duality of conjunctive and disjunctive rules in Section 2.1, it is clear that, if for each X_j both the original variable X_j and the complemented variable $\neg X_j$ are included as predictors, then a conjunctive rule with error rate π is formally equivalent with a disjunctive rule with error $1 - \pi$. The analyses in section 4 for the illustrative example did include for every predictor both the original and the complemented variable and resulted in values for π that are (considerably) smaller than 0.5. Hence, for this particular case a conjunctive rule is found to be more appropriate than a disjunctive one, which is a result established a priori that corresponds with substantive theories on emotion definitions.

As a final comment, we note that the Bayesian (as well as the deterministic) approach to Boolean regression seems to be less useful when the number of observations is very large. For, in general, the variance in $D(y, \theta)$ among θ is expected to increase with the number of observations. In particular, the difference between the best and the second best model most likely increases with the number of observations. From equations(7) and (8), it then follows that the larger the number of observations, the sharper the (marginal) posterior distribution (for θ) is peaked. This implies that if a model θ for some

data set with, say, $n = 1{,}000$ observations has 10% discrepancies and θ^* has 10.01% discrepancies, then θ has a much higher posterior density than θ^*. How this finding can be reconciled with the intuition that both models are about equally likely to be true is one of the objectives for further research.

References

Biswas, N.N. (1975). *Introduction to logic and switching theory.* New York: Gordon and Breach.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gelman, A., Leenen, I., Van Mechelen, I., & De Boeck, P. (1999). *Bridges between deterministic and probabilistic models for binary data.* Manuscript submitted for publication.

Gelman, A., Meng, X.L., & Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Halder, A.K. (1978). Grouping table for the minimization of n-variable Boolean functions. *Proceedings of the Institution of Electric Engineers London*, **125**, 474–482.

Leenen, I., & Van Mechelen, I. (1998). A branch-and-bound algorithm for Boolean regression. In: I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Data Highways and Information Flooding, a Challenge for Classification and Data Analysis* (pp. 164–171). Berlin: Springer–Verlag.

McCluskey, E.J. (1965). *Introduction to the theory of switching circuits.* New York: McGraw–Hill.

McKenzie, D.M., Clarke, D.M., & Low, L.H. (1992). A method of constructing parsimonious diagnostic and screening tests. *International Journal of Methods in Psychiatric Research*, **2**, 71–79.

Mickey, M.R., Mundle, P., & Engelman, L. (1983). Boolean factor analysis. In W.J. Dixon (Ed.), *BMDP statistical software* (pp. 538–545, p. 692). Berkeley, CA: California Press.

Ragin, C.C., Mayer, S.E., & Drass, K.A. (1984). Assessing discrimination: A Boolean approach. *American Sociological Review*, **49**, 221–234.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151–1172.

Sen, M. (1983). Minimization of Boolean functions of any number of variables using decimal labels. *Information Sciences*, **30**, 37–45.

Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy.* San Francisco: Freeman.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327–352.

Van Mechelen, I. (1988). Prediction of a dichotomous criterion variable by means of a logical combination of dichotomous predictors. *Mathémathiques, Informatique et Sciences Humaines*, **102**, 47–54.

Van Mechelen, I., & De Boeck, P. (1990). Projection of a binary criterion into a model of hierarchical classes. *Psychometrika*, **55**, 677–694.

Wierzbicka, A. (1992). Defining emotion concepts. *Cognitive Science*, **16**, 539–581.

Appendix: Deriving posterior distributions

We first compute the prior predictive distribution $p(y)$:

$$\begin{aligned}
p(y) &= \sum_{\vartheta\in\Theta}\left[\int_0^1 p(y|\vartheta,\pi)p(\vartheta)p(\pi)d\pi\right] \\
&= \sum_{\vartheta\in\Theta}\left[\int_0^1 \pi^{D_\vartheta}(1-\pi)^{n-D_\vartheta}\frac{1}{2^k}1\; d\pi\right] \\
&= \sum_{\vartheta\in\Theta}\left[\frac{1}{2^k}\frac{\Gamma(D_\vartheta+1)\Gamma(n-D_\vartheta+1)}{\Gamma(n+2)}\;\times\right. \\
&\qquad\qquad \left.\int_0^1 \frac{\Gamma(n+2)}{\Gamma(D_\vartheta+1)\Gamma(n-D_\vartheta+1)}\pi^{D_\vartheta}(1-\pi)^{n-D_\vartheta}d\pi\right] \\
&= \sum_{\vartheta\in\Theta}\left[\frac{1}{2^k}\frac{D_\vartheta!,(n-D_\vartheta)!}{(n+1)!}\right] \\
&= \frac{1}{2^k(n+1)}\sum_{\vartheta\in\Theta}\frac{1}{\binom{n}{D_\vartheta}}\;,
\end{aligned}$$

the integral in the third step being equal to 1 as it is the area under a Beta density.

For the posterior distribution of (θ,π), we start from Eq. (5):

$$p(\theta,\pi|y) = \frac{p(y|\theta,\pi)p(\theta,\pi)}{p(y)}$$

$$
\begin{aligned}
&= \frac{\pi^{D_\theta}(1-\pi)^{n-D_\theta}\frac{1}{2^k}}{\frac{1}{2^k(n+1)}\sum_{\vartheta\in\Theta}\frac{1}{\binom{n}{D_\vartheta}}} \\
&= \frac{(n+1)\,\pi^{D_\theta}(1-\pi)^{n-D_\theta}}{\sum_{\vartheta\in\Theta}\frac{1}{\binom{n}{D_\vartheta}}}
\end{aligned}
$$

To derive the marginal posterior distribution of θ, π is integrated out in the joint posterior distribution for θ and π in the formula above.

$$
\begin{aligned}
p(\theta|y) &= \int_0^1 p(\theta,\pi|y)\,d\pi \\
&= \int_0^1 \frac{(n+1)\,\pi^{D_\theta}(1-\pi)^{n-D_\theta}}{\sum_{\vartheta\in\Theta}\frac{1}{\binom{n}{D_\vartheta}}}d\pi \\
&= \frac{\frac{1}{\binom{n}{D_\theta}}}{\sum_{\vartheta\in\Theta}\frac{1}{\binom{n}{D_\vartheta}}}\int_0^1 \binom{n}{D_\theta}(n+1)\,\pi^{D_\theta}(1-\pi)^{n-D_\theta}d\pi,
\end{aligned}
$$

the latter integral being 1 as it is again the area under a Beta density.