Bridges between deterministic and probabilistic models for binary data*

Andrew Gelman, Department of Statistics, Columbia University, New York, USA
Iwin Leenen, Psychology Department, Katholieke Universiteit Leuven, Belgium
Iven Van Mechelen, Psychology Department, Katholieke Universiteit Leuven, Belgium
Paul De Boeck, Psychology Department, Katholieke Universiteit Leuven, Belgium
Jeroen Poblome, Sagalassos Archaeology Research Project, Katholieke Universiteit Leuven, Belgium

May 9, 2000

Abstract

For the analysis of binary data, various deterministic models have been proposed, which are generally simpler to fit and easier to understand than probabilistic models. We claim that corresponding to any deterministic model is an implicit stochastic model in which the deterministic model fits imperfectly, with errors occurring at random. In the context of binary data, we consider two error models: in the first model, all predictions are equally likely to be in error; in the second model, the probability of error depends on the model prediction. We show how to fit these models using a stochastic modification of deterministic optimization schemes.

The advantages of fitting the stochastic models explicitly (rather than implicitly, by simply fitting a deterministic model and accepting the occurrence of errors) include quantification of uncertainty in the deterministic model's parameter estimates, better estimation of the true model error rate, and the ability to check the fit of the model nontrivially. We illustrate with a simple theoretical example of item response data and with empirical examples from archaeology and the psychology of choice.

Keywords: binary data, deterministic models, Guttman scaling, implicit stochastic models, Metropolis-Hastings algorithm, parallelogram model, psychometrics

1 Introduction

Data are often fit deterministically; that is, observed outcomes are modeled as being exactly predictable from a small set of parameters. Deterministic models are often easy to understand. However, their conceptual status is somewhat unclear, given the inevitable discrepancies between model and data. Also, overfitting can cause error rates to be understated. Systematic model checking is tricky due to the lack of any sampling distribution with which to compare the data. In statistics, it is standard to use probabilistic models; however these are more complicated and require additional effort to set up.

*We thank David H. Krantz and Marcel Croon for helpful comments. This work was supported in part by fellowship F/96/9 and Grants OT/96/10 and PDM/99/037 of the Research Council of Katholieke Universiteit Leuven and Young Investigator Award DMS-9457824 of the U.S. National Science Foundation.

In this paper, we present a general method of taking any deterministic model for binary data and using it to construct related probabilistic models that allow the possibility of model error. We discuss how to set up such models and perform inference and model checking using Bayesian methods.

The claim of this paper is not that the proposed models are ideal. Rather, we hope to use them to bridge between deterministic and stochastic frameworks. Although deterministic models are not well-analyzed in the statistical literature, they are important data analysis tools in a wide variety of applications. Given that deterministic models are often useful and are commonly used, we would like to do things with them (quantify uncertainty, estimate error probabilities more accurately, check model fit) that are routine with stochastic models. What is new in this paper is the development of a general approach for constructing implicit stochastic generalizations, along with methods for computation and checking the fit of the models to data.

We begin in Section 2 with a general approach to constructing stochastic models for binary data, using any existing deterministic model as a basis for the construction. We set up two stochastic models, one with a single probability of errors, and another with different probabilities of false positive and false negative errors. In Section 3, we show how any existing iterative optimization algorithm can be generalized to a Metropolis-Hastings algorithm for drawing posterior simulations for the stochastic models. We also discuss how to use these simulations to construct random replicated datasets that can be used to check model fit.

In the next two sections, we explore and illustrate these ideas in a number of examples. Our goal in setting up these examples is to work with relatively simple but nontrivial models that have been fit deterministically in applications. In particular, we consider two models in various substantive areas: Guttman scaling and Coombs's parallelogram model. In Section 4, we apply deterministic and stochastic Guttman scaling to two different simulated datasets, with the goal of comparing the abilities of the deterministic and stochastic models to recover true parameter values. In Section 5, we develop the model and algorithm for the stochastic parallelogram model and fit them to two empirical datasets from different substantive areas: archaeology and the study of choice. These data had previously been analyzed using the deterministic model. Here, the point is to see what is gained by generalizing to the stochastic models with one and two error probabilities.

For both real and fake-data examples, the stochastic model has the ability to distinguish between false negative and false positive error rates, which has implications both for inferences about model parameters and for generalizing the model and predicting errors for future data. Most encouragingly, aspects of the data that were *not* fit well by the deterministic models were uncovered using posterior predictive checks applied to their stochastic generalizations. We conclude in Section 6 with a discussion of these and other issues.

2 Models

2.1 Already-existing deterministic model

Consider a data vector $y = (y_1, \ldots, y_n)$ of 0's and 1's, modeled as a specified deterministic function of a vector z of unknown quantities including parameters and potential observables such as latent class memberships. The deterministic model assumes that y can be exactly predicted given z (which is typically of lower dimension than y). In practice, when such models are fit to empirical data, the fit is not perfect, and so it is more appropriate to speak of a vector of predicted data $\hat{y}(z) = (\hat{y}_1(z), \ldots, \hat{y}_n(z))$. We emphasize that we are still working with a fundamentally deterministic model, in that $\hat{y}$ is a vector of 0's and 1's that potentially could be equal to an observed data set. This is in contrast to stochastic models such as logistic regressions in which the predictions $\hat{y}_i(z)$ are continuous-valued and represent $\Pr(y_i = 1)$.

A typical goal in fitting a deterministic model to actual data is to find a z that minimizes the number of errors, that is, the overall number of discrepancies between observed y and predicted $\hat{y}(z)$, which we label as D,

$$D(y, \hat{y}(z)) = \sum_{i=1}^{n} |y_i - \hat{y}_i(z)|, \tag{1}$$

and then to summarize the model by the estimate z and the error rate, D/n.

2.2 Derived stochastic model with one error probability

We can formalize an implicit stochastic model by modeling the outcomes y_i as independent, with

$$\Pr(y_i = \hat{y}_i | z, \pi) = 1 - \pi, \tag{2}$$

where π is the expected error rate corresponding to this model. The likelihood of the data y under this model is

$$p(y|z, \pi) = \pi^{D(y,\hat{y}(z))}(1 - \pi)^{n - D(y,\hat{y}(z))}. \tag{3}$$

Estimating z to minimize the error rate, and then using that as an estimate of the expected error rate, is simply maximum likelihood estimation of (z, π) in (3), or the posterior mode under a uniform prior distribution on (z, π). In the Bayesian context, we use a uniform prior distribution not because we believe it to be most scientifically appropriate for these problems, but because it represents a minimal expansion of the already-existing deterministic model. The expansion from (1) to (2) is analogous to the extrapolation, in continuous models, of the Gaussian error distribution from least squares estimation.

2.3 Derived stochastic model with two error probabilities

An immediate generalization of (2) still specifies independent errors but allows different error probabilities for responses predicted to be 0 or 1:

$$\Pr(y_i = \hat{y}_i(z)|z, \pi_0, \pi_1) = \begin{cases} 1-\pi_0 & \text{if } \hat{y}_i(z) = 0 \\ 1-\pi_1 & \text{if } \hat{y}_i(z) = 1. \end{cases} \tag{4}$$

This is a stochastic version of a deterministic fitting procedure that minimizes some balance of the frequencies of the two kinds of errors (false negatives and false positives), which we label

$$\begin{aligned} D_0(y, \hat{y}(z)) &= \sum_{i|\hat{y}_i(z)=0} |y_i - \hat{y}_i(z)| \\ D_1(y, \hat{y}(z)) &= \sum_{i|\hat{y}_i(z)=1} |y_i - \hat{y}_i(z)|, \end{aligned} \tag{5}$$

with maximum frequencies of occurrence

$$\begin{aligned} n_0(\hat{y}(z)) &= \sum_{i=1}^{n} (1 - \hat{y}_i(z)) \\ n_1(\hat{y}(z)) &= \sum_{i=1}^{n} \hat{y}(z). \end{aligned}$$

The likelihood of the data under this model is

$$p(y|z, \pi_0, \pi_1) = \pi_0^{D_0(y,\hat{y}(z))}(1-\pi_0)^{n_0(\hat{y}(z))-D_0(y,\hat{y}(z))}\pi_1^{D_1(y,\hat{y}(z))}(1-\pi_1)^{n_1(\hat{y}(z))-D_1(y,\hat{y}(z))}. \tag{6}$$

One can, of course, go on to construct more complicated models with dependent error structure, but this would go beyond our purpose of setting up stochastic models that are minimal generalizations of deterministic models.

2.4 Modeling issues

To complete the Bayesian model, we must address two more issues: identifiability of the error parameters and parameterization for the prior distribution.

For many models, one must constrain the error probability π to be less than 1/2 so that the parameters are identified (for example, a prediction of 0 with an 80% error probability is equivalent to a prediction of 1 with a 20% error probability). For the two-error-probability model, one can constrain $\pi_0 + \pi_1$ to be less than 1. In practical applications, this should not be a serious concern, since it would not make sense to use a deterministic model with an error rate even close to 1/2.

As discussed at the end of Section 2.2, deterministic optimization can be identified as maximum likelihood, and so a uniform prior distribution would seem logical for a minimal Bayesian generalization. This still does not determine the model, however, until a parameterization is chosen. As with statistical transformations in general, we cannot give general advice except that it should be reasonable. The examples in the later sections of this paper give some sense of the choices involved in setting up this part of the model.

2.5 Advantages of the stochastic models

In addition to the conceptual advantage of explicitly admitting that the model can have errors, we see three principal benefits of performing statistical inference with model (2) or (4) instead of optimization with the deterministic model. First, the stochastic formulations allow for uncertainty in z (and also π, but that typically is of less interest), which is particularly relevant when many different possible z's yield predicted data $\hat{y}(z)$ that closely fit the observed y. Second, a stochastic model takes account of mispredictions and should give a better estimate of the true model error than is obtained by the deterministic estimate, which underestimates model error because it optimizes the fit[1]. Third, a stochastic model defines a predictive distribution for future data, y^{rep}. One can use this predictive distribution as a reference distribution for checking the fit of observed data y to the model (see the end of Section 3 for details). In contrast, the deterministic model can only be tested trivially: any mispredictions indicate a model failure.

3 Computation and model checking

As discussed in Section 2, we assume that some sort of deterministic procedure is already available for estimating the vector of unknowns z from the observed data y, and we wish to generalize to draw simulations from the posterior distribution of (z, π) or (z, π_0, π_1), and then from the predictive distribution of future data y^{rep}. Inference is summarized by the simulations of z, error is summarized by discrepancies between data y and predictions $\hat{y}(z)$, and models can be checked by comparing observed data y to replicated data sets y^{rep}.

3.1 Computing the model with one error probability

For the single-error-rate model, we proceed in the following general steps:

1. Run the fitting procedure for the deterministic model from several different starting points (or run several different fitting procedures) to get several estimates of z. Use relatively simple fitting algorithms if possible, since these estimates are intended to be used merely as starting points for the simulations.

[1]If the true parameter vector z^{true} were unknown, then the true error rate of the observed data, $D(y, \hat{y}(z^{\text{true}}))$, would be an unbiased estimator of the error probability π. The deterministic estimating procedure chooses the parameters that minimize the error rate. For a nondegenerate model, $\mathrm{E}(\min_z(D(y, \hat{y}(z)))) < \mathrm{E}(D(y, \hat{y}(z^{\text{true}})))$, and so the error rate of the observed data with respect to the best-fitting model is a negatively-biased estimate of the true error rate.

2. For each estimated z, obtain a starting estimate of π at the adjusted error rate $(D(y, \hat{y}(z)) + 1)/(n+2)$. (We add 1 in the numerator and 2 in the denominator to avoid the possibility of an initial estimate of $\pi = 0$ or 1, which would interfere with the later updating steps.)

3. Run several parallel sequences of a Metropolis-Hastings algorithm (see, e.g., Gilks, Richardson, and Spiegelhalter, 1996) using the above estimates of (z, π) as starting points $(z^{(0)}, \pi^{(0)})$. At each iteration $t = 1, 2, \ldots$, perform the following steps.

   (a) Draw a candidate value z^* from a jumping distribution $J(z^*|z^{(t-1)})$ conditional on the current value $z^{(t-1)}$. (That is, $J(z^*|z^{(t-1)})$ is defined as the probability of considering the candidate value z^*, conditional on the current value of z at this step in the algorithm.)

   (b) Compute the importance ratio
$$r = \frac{p(y|z^*, \pi^{(t-1)})}{p(y|z^{(t-1)}, \pi^{(t-1)})} \frac{J(z^{(t-1)}|z^*)}{J(z^*|z^{(t-1)})}, \tag{7}$$
   with $p(y|z, \pi)$ given by (3).

   (c) Accept or reject the candidate value. That is, set
$$z^{(t)} = \begin{cases} z^* & \text{with probability } \min(1, r) \\ z^{(t-1)} & \text{otherwise.} \end{cases}$$

   (d) Update π by drawing from a Beta$(D(y, \hat{y}(z^{(t)})) + 1, n - D(y, \hat{y}(z^{(t)})) + 1)$ distribution, if necessary with the constraint that $\pi < 1/2$ (which can be achieved by resampling from the beta distribution until a value less than 1/2 has been drawn).

   Repeat these steps independently for the several parallel simulations until the sequences appear mixed (see, e.g., Gelman and Rubin, 1992).

4. Select L (for example, 1000) random draws of (z, π) from the simulated sequences to represent posterior simulation draws.

5. From each posterior simulation draw (z_l, π_l), simulate a replicated data set $y^{\text{rep}\, l}$: first, compute $\hat{y}(z^l)$; second, simulate the components of $y^{\text{rep}\, l}$ from $\hat{y}(z^l)$ independently using (2).

3.2 Computing the model with two error probabilities

Computation for the two-error-probability model is identical except for the following changes. In step 2, initial estimates are required for both π_0 and π_1, for which one can use $(D_0(y, \hat{y}(z)) + 1)/(n_0(\hat{y}(z)) + 2)$ and $(D_1(y, \hat{y}(z)) + 1)/(n_1(\hat{y}(z)) + 2)$, respectively. In step 3b, equation (7), use $p(y|z, \pi_0, \pi_1)$ from (6) in place of $p(y|z, \pi)$. Finally, break step 3d into two steps: for $k = 0, 1$, draw π_k from the Beta$(D_k(y, \hat{y}(z^{(t)})) + 1, n_k(\hat{y}(z)) - D_k(y, \hat{y}(z^{(t)})) + 1)$ distribution (if necessary, repeating *both* steps until $\pi_0 + \pi_1 < 1$).

3.3 Computational details

For either of the models, given that step 1 has already been programmed, the most difficult remaining task above is step 3a: specifying and drawing from a jumping distribution, $J(z^*|z^{(t-1)})$. The specific algorithm here necessarily depends on the model being fit, but we can give two general comments here. First, the fitting algorithm for the deterministic model may well provide a nucleus for the Metropolis-Hastings algorithm. For example, if the deterministic algorithm is an iterative maximization ("greedy") algorithm, then one can jump by taking a single maximization step and adding to it a random component. If it is convenient to update components of z separately, then steps 3(a,b,c) can be divided into sub-iterations, as in a Metropolis-within-Gibbs algorithm. We illustrate with the examples in Sections 4 and 5. Our second general comment is that, even if the Metropolis-Hastings algorithm does not converge (in that the different simulated sequences do not mix), the result of the above steps may be considered an improvement upon the deterministic point estimates, which we use as starting points.

3.4 Model checking

The algorithm above yields a set of simulation draws $(z^l, \pi^l, y^{\text{rep}\,l})$, $l = 1, \ldots, L$. We can use these draws to check the model fit using posterior predictive checks (see Rubin, 1984, Gelman, Carlin, Stern, and Rubin, 1995, and Gelman, Meng, and Stern, 1996) as follows. First, define a test summary $T(y, z)$ to summarize some aspect of interest of the data or its discrepancy from the model. Second, compute the realized values of the test summary, $T(y, z^l)$ and the replicated values $T(y^{\text{rep}\,l}, z^l)$ under the implicit stochastic model. Third, compare the realized to the replicated values, with statistical significance measured by the posterior predictive p-value, $\Pr(T(y^{\text{rep}\,l}, z^l) > T(y, z^l))$, which can be estimated by the proportion of the L simulations for which $T(y^{\text{rep}\,l}, z^l) > T(y, z^l)$. Different test summaries can be used to check the fit of different aspects of the data. Often, the most effective test summaries are simple graphical displays of the data and model fit; other times, it is helpful to check significance using p-values. We illustrate both kinds of checking in the examples.

4 Theoretical example

4.1 Guttman scale model and stochastic extensions

As an illustrative example, we consider the deterministic scalogram or Guttman scale model (Guttman, 1950, Torgerson, 1962, Coombs, 1964). The data analyzed are a matrix (y_{ij}) of binary responses of individuals $i = 1, \ldots, I$ on items $j = 1, \ldots, J$, where $y_{ij} = 1$ or 0 corresponds to individual i giving a correct or wrong answer, respectively, for item j. According to the Guttman scale model, each individual is assumed to have a continuous ability parameter u_i and each item to have a continuous difficulty level v_j, with $y_{ij} = 1$ if and only if $u_i > v_j$. (Because the parameters are continuous, we rule out the possibility of any two elements of z being exactly equal.) This model, being an archetypal deterministic model in psychometrics and sociometrics, is used here to illustrate our general ideas.

During the past few decades various stochastic extensions of the Guttman scale have been proposed, most notably the Rasch (logistic regression) model (see, e.g., Fischer and Molenaar, 1995). In addition to the fully probabilistic Rasch model, other stochastic extensions have been advanced that preserve the original deterministic scalogram model as the core of the stochastic model, similar to the approach taken in the present paper. Examples include the latent class models proposed by Goodman (1975) and Proctor (1970). In particular, Goodman's (1975) model includes latent classes of "scalable" persons whose response pattern exactly corresponds to an admissible pattern of a given Guttman scale, in addition to a single class of "nonscalable" persons within which independence of the items is assumed. Proctor's (1970) model includes latent classes that each correspond to a true Guttman response pattern, and further assumes within each class a Bernoulli type of error process (governed by one or two error probabilities), similar to the one proposed in the present paper. Other authors implicitly advocate stochastic extensions of the Guttman scale model based on a Bernoulli type of error process when proposing various goodness-of-fit tests for the model (see, e.g., Goodman, 1959; Hoffman, 1979). A major difference between the extended scalogram model proposed in the present paper and the mixed deterministic/stochastic models just mentioned is that all of the latter assume that a true Guttman order of the items is known. One of the distinct advantages of our approach therefore is the posterior uncertainty about the underlying item order it may reveal.

4.2 Hypothetical data

We fit the deterministic and stochastic Guttman models to two hypothetical datasets, one constructed to fit the stochastic Guttman model fairly well, and the other based on the Rasch model, which differs quite a bit from Guttman scaling.

4.2.1 Hypothetical data simulated under an approximate Guttman model

We consider a hypothetical data set with $I = 40$ individuals and $J = 20$ items, with the u_i's and v_j's ranging from 1 to 60, and with the item difficulties spread evenly through the abilities. Our data set is created by first computing the predicted responses and then randomly switching 25% of the predicted 0's to 1's (to simulate guessing on a multiple-choice test). Next we switch some of the predicted 1's to 0's in a slightly more complicated fashion: For the questions that are easier than average, we randomly switch 5% of the predicted 1's to 0's; for the questions that are harder than average, we randomly switch 15% of the predicted 1's to 0's (so that mistakes are more likely on difficult questions). Most of the predicted 1's come from easier questions, and the total error rate for the predicted 1's is 7.5%. The error rate is 25% of the predicted 0's and 16.2% of all the data. The left side of Figure 1 displays the data, with the switched responses shown in boldface.

These data were created to simulate the following conditions. First, the deterministic model describes the data fairly well (a true error rate of less than 1/6), which seems reasonable since one should probably not even consider a deterministic model if the error rate is much higher. Second, the error rates differ for predicted 0's and 1's, which seems realistic and is also of interest for comparing our two error models. Third, error probabilities vary slightly with difficulty, and as a result the implicit assumptions underlying the deterministic model are reasonable but not exactly valid.

4.2.2 Hypothetical data simulated under a Rasch model

As a comparison, we also fit the Guttman model to data for which that model is not appropriate. The right side of Figure 1 shows this data set, which was simulated under a Rasch model, with abilities u_i and difficulties v_j as in the Guttman model, but then assigning 1's to cells (i, j) independently, with probabilities $\Pr(y_{ij} = 1) = \text{logit}^{-1}((u_i - v_j)/\sigma)$. We set $\sigma = 10$, which yielded an aggregate expected error rate of 17%, close to that of the simulated example in Section 4.2.1. For either model, we define an "error" as occurring when $y_{ij} = 1$ when $u_i < v_j$, or $y_{ij} = 0$ when $u_i > v_j$. Due to the symmetry in our setup, the expected error rates for the predicted 0's and 1's are the same here. The actual simulated data have error rates of 17.5% and 16.8% for predicted 0's and 1's, respectively.

We created this second simulated data set in order to see how well the deterministic and stochastic Guttman models would work in fitting ordered data that come from a model that deviates further from the stochastic Guttman model. In particular, the mispredictions of the Rasch model are not uniformly distributed in the data but rather occur more often when estimated probabilities are near 0.5, which occurs around the diagonal of the table, where ability and difficulty parameters are close to each other. Given a nonnegligible occurrence of probabilities near 0.5 (as in this example), fitted deterministic and stochastic Guttman models can be predicted to yield underestimates of the true error rate, as the models will be seduced to come too close to the data (at the expense of the true underlying ability/difficulty order).

4.3 Deterministic model

We fit the same Guttman model to both simulated data sets. In the notation of Section 2.1, the predicted data are defined by $\hat{y}_{ij}(z) = 1_{u_i > v_j}$, and the unknown parameters are $z = (u, v) = (u_1, \ldots, u_I; v_1, \ldots, v_J)$.

We fit the deterministic model to data y by iteratively updating u and v, as follows.

1. Start by ordering the components of u based on the number of correct responses in the data: the individual i with the lowest value for $\sum_{j=1}^{J} y_{ij}$ is assigned ability $u_i = 1$, the individual with the next-lowest value is assigned $u_i = 2$, and so forth. A random number generator is used to break any ties.

2. Set $v = (v_1, \ldots, v_J)$ so as to minimize the error rate $D(y, \hat{y}(u,v))$ (see (1)), conditional on the current estimate of u. This can be done straightforwardly by updating the difficulty parameters v_j one at a time to minimize the error rate for the responses to each question. Once again, use a random number generator to break any ties.

3. Similarly, set u to minimize the error rate conditional on the current estimate of v.

4. Repeat steps 2 and 3 until the error rate no longer decreases.

Because of the random elements in this algorithm, it can be repeated several times to get slightly different answers.

4.4 Stochastic models

In setting up the stochastic models, the only tricky point is assigning a prior distribution to the continuous parameters (u, v). We choose to assign equal prior probabilities to all the $(I + J)!$ orderings of the ability and difficulty parameters. When performing the computations, we identify the model by the rank ordering of the $I + J$ parameters. (In the Guttman model, no information is conveyed in the parameters beyond their ranking.) Because one can switch all the predicted 0's and 1's by simply reversing the ranks, we constrain the error probabilities as described in Section 3.3.

We fit the stochastic models using the algorithm in Section 3, performing step 3 (updating z) in several sub-iterations as follows:

3. For $k = 1, \ldots, I + J - 1$:

   (a) Create z^* by switching the kth and $(k+1)$st ordered values of z. We now consider two possibilities:

      i. If the two parameters being switched are both ability parameters or both difficulty parameters, then switching them has no effect on the model predictions and thus on the likelihood, so $r = 1$ in (7) and the switch is automatically accepted. (The factor of $J(z^{(t-1)}|z^*)/J(z^*|z^{(t-1)})$ disappears in (7) because both jumping probabilities are equal to 1 in this jumping rule.)

ii. If one of the parameters being switched is an ability parameter u_i and the other is an item parameter v_j, then switching them affects the likelihood only through the data point y_{ij}. If we let $h = \hat{y}_{ij}(z^{(t-1)})$, then the ratio r in (7) for this jump is simply $\pi_{1-h}^{(t-1)}/(1-\pi_h^{(t-1)})$ if $y_{ij} = h$ or $(1-\pi_{1-h}^{(t-1)})/\pi_h^{(t-1)}$ otherwise. (For the one-error-probability model, both π_h and π_{1-h} are simply π.)

4.5 Results

4.5.1 Fitting the deterministic model

We begin our computations by fitting the deterministic model to the data sets in Figure 1, for each example running the algorithm five times to obtain five different solutions. (These data sets are large enough that the random starting points and random tiebreaking in the algorithm force them to converge to different points upon replication.) For each example, Table 1 shows the error rates $D_0(y, \hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$ for the estimated parameters z from each of the 5 runs of the optimization algorithm; as could be expected, they all underestimate the true total error rate (although, interestingly, the error rate for predicted 1's is overestimated for the approximate Guttman data). As predicted, the error rates are drastically underestimated when the deterministic (all-or-none) Guttman model is fit to data generated under the (continuous) Rasch model.

4.5.2 Fitting the stochastic models

For each of the two examples, we then use the five deterministic solutions as starting points for independent runs of the Metropolis-Hastings algorithms for the one-error-probability model and, separately, for the two-error-probability model. For each of these models, we find that 2000 simulations were sufficient for approximate convergence, in the sense that the potential scale reduction factor is less than 1.2 for all of the components of (u, v), all of the variance components, and the log-likelihood (see Gelman and Rubin, 1992). We discard the first half of the simulations and save every fifth iteration (to save computer storage). For each of the two models fit to each of the two data sets, we then have 1000 draws of the vector $z^l = (u, v)^l$ from the posterior distribution; we use these for all further computations.

Table 2 shows the estimated error rates and error probabilities for five random draws from the posterior distribution for the one- and two-error-probability models as fit to the two simulated data sets.

4.5.3 Inference and model checking for the approximate Guttman data

We begin by examining the estimated error rates for the model with one error probability fit to the simulated approximate Guttman data. Figure 2a shows a histogram of the posterior distribution of the total error rate $D(y, \hat{y}(z^l))/n$ from 1000 draws l from the posterior distribution. As would be hoped, the estimated total error rate is close to the true error rate of 0.162, with a wide range of uncertainty due to the relatively small sample size and large number of parameters in the model. We also examine separately the conditional error rates $D_0(y, \hat{y}(z^l))/n_0(\hat{y}(z^l))$ and $D_1(y, \hat{y}(z^l))/n_1(\hat{y}(z^l))$, as displayed in Figure 2b. The rate of the more common false negative error (predicted 0's, observed 1's) is underestimated and the rate of the less common false positive error is overestimated.

Using the stochastic formulation, we can uncover this misfit of the model using posterior predictive simulations. A notable feature of Figure 2b is that the error *rates* for predicted 0's and 1's differ greatly, even while the error *probabilities* are constrained by the model to be equal. This would seem to indicate a model violation. We can check this with a posterior predictive check, using the difference in error rates as a test summary, $T(y, z) = D_0(y, \hat{y}(z))/n_0(\hat{y}(z)) - D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$, and then comparing the realized values $T(y, z^l)$ to the replicated values $T(y^{\text{rep}\,l}, z^l)$ as described in Section 3.4. In our simulations, the mean value of the realized test summary $T(y, z^l)$ is 0.11, the mean value of the replicated test summaries $T(y^{\text{rep}\,l}, z^l)$ is 0.00 (to two significant figures), and in all the simulations, $T(y, z^l)$ had a larger value than $T(y^{\text{rep}\,l}, z^l)$. This clearly indicates a lack of fit of the single error-probability model.

We are then lead naturally to the model that allows the error probabilities to differ for predicted 0's and 1's. Figure 3 displays the estimated error rates when fitting this model to the approximate Guttman data. This time, when we check $T(y, z) = D_0(y, \hat{y}(z))/n_0(\hat{y}(z)) - D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$, we find a good fit: the average values of $T(y, z^l)$ and $T(y^{\text{rep}\,l}, z^l)$ are both 0.21, and $T(y, z^l)$ exceeds $T(y^{\text{rep}\,l}, z^l)$ in 50.4% of the simulations. That is, the observed data fits in well among the replicated datasets. It is no surprise that this model fits this aspect of the data since this fitting is directly addressed by the two error probabilities.

Of course, the actual data do *not* come from the model in this case. (Recall from Section 4.2.1 that the data were simulated to have more prediction errors for difficult items, which is why we refer to these as the "approximate" stochastic Guttman data.) So all we have learned is that the differential error probabilities can be reasonably estimated, not that the model "fits" in some general sense. As always, the criterion of fit of a model depends on the uses to which it might be put. For example, we can further check the model by considering, as a test summary, the correlation between errors and difficulty across the 20 items. From the posterior simulations, the expected value of this test summary in the realized data is 0.31 and the expected value in the replicated data is 0.14, and in 99.3% of the simulations, the realized value exceeds the replicated value, thus indicating a poor fit to this aspect of the data.

4.5.4 Inference and model checking for the Rasch data

We similarly fit the stochastic Guttman models with one and two error probabilities to the simulated Rasch data pictured on the right side of Figure 1. For brevity, we skip straight to the two-error-probability model in our exposition. Figure 4 displays the posterior distribution of estimated total and conditional error rates, by analogy to Figure 3. As predicted, the stochastic Guttman model dramatically underestimates the error rates, indicating that it is overfitting the data (although not as badly as with the deterministic version of the model).

This is a serious model misfit, but of course in practice we cannot check a model by comparing its posterior distribution to the true parameter values. Rather, we check the fit by comparing test summaries to their distribution under the model. In this case, the most straightforward check is to use, as a test summary, the correlation between errors $|y_{ij} - \hat{y}_{ij}(z)|$ and the absolute value of the difference between ability and item difficulty $|u_i - v_j|$. Under the stochastic Guttman model, we would expect no particular correlation here, except what might arise as an artifact of the two different error probabilities. In data on tests of continuous abilities, however, one might reasonably expect prediction of y_{ij} to be less reliable when the difficulty of item j is close to the ability of individual i. This would induce a negative correlation between $|y_{ij} - \hat{y}_{ij}|$ and $|u_i - v_j|$.

As a matter of fact, the expected value of this test summary is -0.23 for the realized data but only 0.02 for the replicated data, and the realized value is lower than the replicated in all the simulations, thus revealing this aspect of the data to be inconsistent with the model.

5 Applied examples using Coombs's parallelogram model

We now consider a Bayesian extension of the so-called *parallelogram model* initially presented by Coombs (1964) for the analysis of choice data in psychology. The parallelogram model has been widely used and studied in various other domains, though under different names: In mathematics, Fulkerson and Gross (1965) studied the consecutive ones property for rows in a binary matrix; similar concepts have been used in archaeology for seriation (see Kendall, 1969, Goldmann, 1971, and Arabie and Hubert, 1992) and in computer science for information retrieval (Ghosh, 1972). Fully probabilistic extensions of the parallelogram model have also been developed (see, e.g., Andrich, 1995, and Hoijtink, 1990).

In this section, we first formally present the deterministic parallelogram model (Section 5.1) and its Bayesian extension (Section 5.2). We then illustrate with applications on archaeological data (Section 5.3) and on choice data (Section 5.4).

5.1 Deterministic model

Assume a matrix (y_{ij}) with rows $i = 1, \ldots, I$ and columns $j = 1, \ldots, J$. This matrix represents a parallelogram relation (Coombs, 1964) if each column j can be given a position q_j on an underlying dimension and each row i can be characterized by an ideal point c_i on the dimension and a latitude of acceptance ϵ_i, such that,

$$y_{ij} = \begin{cases} 1 & \text{if } |c_i - q_j| < \epsilon_i \\ 0 & \text{otherwise.} \end{cases}$$

For convenience, we use a reparametrization of Coombs's model (see Leenen and Van Mechelen, 1999), in which each column j is still given a position q_j, but each row is now characterized by an interval $[a_i, b_i]$. The predicted data are then defined by,

$$\hat{y}_{ij}(z) = \begin{cases} 1 & \text{if } a_i \leq q_j \leq b_i \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

with unknown parameters $z = (q, a, b) = (q_1, \ldots, q_J, a_1, \ldots, a_I, b_1, \ldots, b_I)$. As only the ordering of the elements in z is relevant, we add the restriction that $q_j, a_i, b_i \in \{1, \ldots, 2I + J\}$ and that no two elements of z are equal. The parallelogram model suffers from an indeterminacy due to possible reversing of the underlying scale: for any given model z, its reversed counterpart z' (with $z'_h = 2I + J + 1 - z_h$; $h = 1, \ldots, 2I + J$) fits the data equally well.

Leenen and Van Mechelen (1999) have recently proposed an alternating least squares algorithm for fitting the deterministic parallelogram model to a given data set, which somewhat resembles the method proposed in Section 4.3 for fitting the deterministic Guttman scale model:

1. An initial ordering for the column parameters is obtained either randomly or by a built-in heuristic in the algorithm. (In case of a random initial start, one typically repeats the procedure several times.)

2. Optimal estimates for a and b are found conditionally upon y and q, optimality implying a minimal value for the error rate $D(y, \hat{y}(z))$. The contribution of row i to the total error rate can be separated from the contribution of the other rows, and so the intervals for the rows can be updated successively using fast separate computations. Optimal estimates a_i and b_i for row i conditionally upon y and q can be computed using a branch-and-bound algorithm.

3. Analogously, optimal estimates for q are found conditionally upon y, a, b.

4. Steps 2 and 3 are repeated until the error rate no longer decreases.

5.2 Stochastic model

We assign a uniform prior distribution for $z = (q, a, b)$: equal prior probabilities are assigned to all orderings of the row and column parameters that satisfy $a_i < b_i$ for each row i. (There are

$(2I+J)!/2^I$ such orderings.) Although for parallelogram models, unlike for Guttman scale models, it is usually not possible for a given z to find a z' such as to switch all the predicted 0's and 1's, we keep constraining π to be smaller than 0.5.

We fit the stochastic parallelogram model using the algorithm of Section 3 with the following specifications for steps 1 and 3 of the algorithm:

1. We run the deterministic fitting procedure using a random initial start for the column parameters to get several estimates of z. Optionally, each single z may be the best solution retained after several runs with a random initial start.

3. At each iteration $t = 1, 2, \ldots$, we perform the following steps:

   A z^* is created as follows: Initially, if the first ordered value in $z^{(t-1)}$ is a column, it is moved to the right end of the scale. Furthermore, the kth and $(k+1)$st ordered values of $z^{(t-1)}$ $(k = 1, \ldots, 2I+J-1)$ are switched, similarly to step 3(a) in the algorithm for fitting the Guttman scale model in Section 4.4. The following cases can occur:

   - If the two parameters being switched are both row parameters (a_i or b_i) or both column parameters, (q_j), then the switch does not affect the model predictions or the likelihood, so $r = 1$ in equation (7), and the switch is automatically accepted. An exception is made if the parameters pertain to the same row, that is, if they are the lower limit a_i and the upper limit b_i of the same row's interval. If this occurs, the switch is never accepted, since (8) implies that $a_i \leq b_i$.
   - The other possibility is that one of the parameters being switched is a row parameter a_i or b_i, and the other a column parameter q_j. Then, a switch affects the likelihood only through the data point y_{ij}; if we let $h = \hat{y}_{ij}\left(z^{(t-1)}\right)$, it then holds that $r = \pi_{1-h}^{(t-1)}/(1-\pi_h^{(t-1)})$ if $y_{ij} = h$ or $r = (1-\pi_{1-h}^{(t-1)})/\pi_h^{(t-1)}$ otherwise. For the one-error-probability model, both π_h and π_{1-h} are simply π.

Before evaluating convergence of the simulated sequences, each of the models $z^{(t)}$ $(t = 1, 2, \ldots)$ is transformed in two respects: First, scales are reversed when necessary, such that the directions of the scales in all sequences are the same. One may decide on a common direction of the scales by fixing the order of a particular pair of columns, the columns within the selected pair preferably having stable and maximally divergent positions on the scale. To find such a pair, a pre-analysis of the data may be helpful: for example, one may consider the pair of columns j and j' in the data y that has maximal value on $\min(n_{10}, n_{01})$, where $n_{hh'}$ is the number of rows i for which $y_{ij} = h$ and $y_{ij'} = h'$. If more than one pair takes a maximal value on $\min(n_{10}, n_{01})$, one pair can be arbitrarily selected.

Second, scales are condensed by maximally reducing the number of positions on the scale (i.e., each parameter is assigned a new value from $\{1, \ldots, P\}$, where P is minimal under the restriction that the values predicted by (8) are identical). The latter is done to eliminate irrelevant differences between parameter values.

5.3 Application to seriation in archaeology

As a first example, we consider an archaeological study of different ceramic artifacts discovered at several sites within a small geographic area. If the findings at any given site are localized in time, and if any given category of artifact is in use for a contiguous time period, then it is reasonable to fit a parallelogram model to the data, which would yield a time ordering of sites and a time interval for each category of artifact.

We consider data collected at the ancient site of Sagalassos in south-west Turkey, which has been investigated since 1986 by an interdisciplinary team of the Katholieke Universiteit Leuven (Waelkens et al., 1997). The discovery in 1987 of a local potters' quarter was unexpected (Mitchell and Waelkens, 1988) and led to further exploration. Over about six hectares, located to the east of the ancient town, dumps of misfired ceramics are still noticeable at the surface. Architectural ceramics, figurines and oil lamps, cooking and storage vessels, as well as a series of tableware were produced locally with five different clay fabrics (Poblome et al., 1997; Degeest et al., 1997). In economic terms, the newly discovered tableware or Sagalassos red slip ware can be considered the most important feature of this production center. After a Hellenistic antecedent, mass-production of this new type of eastern sigillata started during the Augustan period and lasted into the first half of the seventh century A.D. The ware has been traded intensively in Anatolia, and could already be identified at a series of sites in the eastern Mediterranean and beyond, as far away as Italy and Nubia (Poblome, 1996).

During the Imperial period, only five mass production centers, including Sagalassos, have been identified, supplying the innumerable towns and villages in the Roman east. The different mass production centers or manufactories represented therefore an enormous marketing potential. In this way, the reconstruction of the production organization and trade mechanisms associated with these pre-industrial artisanal units may provide a valuable key in understanding the potential, flexibility and shortcomings of the ancient economy. Due to particular circumstances Sagalassos is actually the only identified production center that is still accessible for archaeological research. A strict methodological framework was therefore developed to examine the local potters' quarter and its products. In this context, only the typo-chronological determination of Sagalassos red slip ware is of importance, however. A total of 78 types and variants of cups, bowls, dishes, plates, and containers were distinguished; after combining similar types and variants, these represented 37 distinct categories of artifacts. The artifacts were found at 36 assemblages, each at a different location or site. We use data from 18 of these sites in our analysis.

In a parallelogram analysis on these data, an interval is assigned to each of the types and a position to each of the assemblages on an underlying dimension, which presumably will turn out to be a time dimension. Thus, from a substantive point of view, a parallelogram analysis can be expected to result in (1) a relative chronological ordering of the assemblages and (2) for each of the ceramic types and variants (further also called artifacts) a time period that indicates when it was actually in use.

The 37×18 data matrix is analyzed both with the deterministic alternating least-squares procedure (Leenen and Van Mechelen, 1999) and the Bayesian procedure presented in Section 5.2 above. Figure 5 shows the data compared to model with the lowest overall number of discrepancies as found by the deterministic algorithm. This overall error rate equals 6.2% (the error rates for predicted 0's and predicted 1's being 8.2% and 5.2%, respectively).

With respect to the Bayesian analysis, we discuss, for brevity, only the results of the two-error-probability model. The Metropolis-Hastings algorithm is executed with four sequences, the starting point of each sequence being a minimally discrepant model out of 100 models obtained by a random start of the deterministic algorithm. The Metropolis-Hastings algorithm converges after almost 50,000 iterations in each sequence, the potential scale reduction factor (Gelman and Rubin, 1992), computed on the last half of the iterations, being below 1.1 for all of the components in (a, q, b) and the error rates π_0 and π_1. We discarded the first half of the simulations and, for computational convenience, saved every tenth draw, resulting in about 10,000 draws z^l from the posterior distribution. Figure 6 shows the posterior distributions of the overall error rate and the error rates for predicted 0's and 1's; the estimated error rate for predicted 0's is much higher than in the best deterministic fit.

Figure 7 displays the data from the perspective of the best-fitting deterministic model. Figure 8 shows similar displays of the observed data with respect to three random draws from the posterior distribution of the model parameters, with the dots in each figure indicating model mispredictions. The error rate for predicted 0's is much higher with the stochastic than the deterministic model, as can be seen clearly by comparing Figures 7 and 8. In models drawn from the posterior distribution, errors near the beginning and the end of an artifact's interval seem to be relatively more likely, whereas in the model yielded by the deterministic algorithm, those errors do not occur at all, as such errors are easily rectified by resizing the artifact's interval in the model. This finding illustrates that error rates in the deterministic model are likely to be underestimated.

To check the hypothesis that errors at the boundaries of the artifacts' intervals are more likely to occur, one may use a posterior predictive check based on a test summary which is sensitive to this type of model violation (for example, the correlation between errors $|y_{ij} - \hat{y}_{ij}(z)|$ and the minimum absolute difference between q_j and either of a_i and b_i, $\min(|q_j - a_i|, |q_j - b_i|)$, similar to the test summary developed in Section 4.5.4).

The chronological evolution of the tableware was independently determined by seriating these ceramic assemblages from specific stratigraphical units and employing internal and external criteria to actually date the relative sequence (Poblome, 1999). Inspection of the models shows that in most cases the parallelogram respects the main lines of the expected chronological sequence of the assemblages and the constituent role the types and variants play. The selection of assemblages contains, for instance, two series of stratigraphically superimposed units (EoN 4-8 and EoN 11-18; L 10N-16N, L 8N-9N, L 5N-7N and L 3N-4N) with, as a consequence, logically fixed chronological relationships within each series, which are found in most of the models. Parallelogram models may therefore be regarded as a valuable first approach for detecting a specific chronological sequence. The interpretative role of archaeology should still be brought into play, however: The presented model does not take into account the intrinsic differences between functionally different contexts, such as floor levels or destruction layers, with possibly different time ranges of similar types and variants. Another area of error may be the ever-present and very site-specific problem of residuality and intrusiveness of archaeological artifacts, whereby types and variants are found in contexts that predate or postdate the artifact. As a result, the types and variants that are more sensitive to residuality and intrusiveness (in particular, 1A100, 1B150, 1B160-3, 1B170, 1B190, 1C170-1, 1F150) are found to have very wide (if not full) intervals in most models. However, that some types and variants are more likely to suffer from residuality and intrusiveness than others is a violation of the Bayesian parallelogram model in that it is assumed that the error rates are identical across variants and types.

To statistically check the validity of the latter assumption, we define test summaries $T_0(y, z)$ and $T_1(y, z)$ as the standard deviations across the artifacts' error rates p_{0i} and p_{1i}:

$$p_{0i} = D_0(y_i, \hat{y}_i(z))/n_0(\hat{y}_i(z)), \qquad \&p_{1i} = D_1(y_i, \hat{y}_i(z))/n_1(\hat{y}_i(z))$$
$$T_0(y, z) = \mathrm{sd}_i p_{0i}, \qquad \&T_1(y, z) = \mathrm{sd}_i p_{1i},$$

where the arrays y_i and $\hat{y}_i(z)$ are, respectively, the data vector and predicted data vector of artifact i. As the model considers artifacts to be homogeneous with respect to their error rate, the expected values for p_{0i} and p_{1i} are equal across artifacts (and equal to π_0 and π_1, respectively).

The test summaries for the realized values $T_b(y, z^l)$ and the replicated values $T_b(y^{\mathrm{rep}\,l}, z^l)$ ($b =$

$0, 1)$ were computed and subsequently compared as discussed in Section 3.4. The mean value across simulations of $T_1(y, z^l)$ is 0.20 and of $T_1(y^{\text{rep}\,l}, z^l)$ is 0.07; in all of the simulations the realized value of T_1 is higher than the replicated value, which indicates that for some artifacts the probability of a false positive is larger than for others, indeed.

Similarly, with respect to the false negative error rates, the mean value of $T_0(y, z^l)$ is 0.23, compared to a mean value of 0.17 in the replications, $T_0(y^{\text{rep}\,l}, z^l)$; in 92% of the simulations the realized value of T_0 was larger than the replicated value. Hence, there is some indication that variants and types differ more than expected under the model with respect to their negative error rates as well.

5.4 Application to data on individual choices

As another example, we reanalyze a subject by object choice data set previously described by Van Mechelen and Van Damme (1994). In this study, 26 second-year psychology students were asked to select, from a set of 25 index cards with uniformed room descriptions from the Housing Service of the University of Leuven, the set of rooms that they would decide to visit when looking for a student room.

The resulting 25×26 binary matrix (y_{ij}) (with $y_{ij} = 1$ if student i selects room j and $y_{ij} = 0$ otherwise) is given in Figure 9. Both the deterministic algorithm (Leenen and Van Mechelen, 1999) and the Bayesian procedure described in Section 5.2 were used to analyze these data. The best-fitting deterministic model had an overall error rate of 11% with the error rate for predicted 0's being 13% and the error rate for predicted 1's being 8%.

Again, we jump straight to results for the two-error-probability model. The model was computed using three sequences in the Metropolis-Hastings algorithm, each with a starting point that was the best resulting from 100 random runs of the deterministic algorithm. The Metropolis-Hastings algorithm converged after 300,000 iterations in each sequence, in that the potential scale reduction factor (Gelman and Rubin, 1992), computed on the last half of the iterations, was below 1.2 for all 79 parameters. We saved every hundredth draw of the last half of the simulations, which yields 4,500 draws z^l from the posterior distribution. Figure 10 shows the posterior distributions of the overall error rate and the error rates for predicted 0's and 1's and Figure 13 shows the data in relation to three draws from the posterior distribution of the orderings.

The next step is to check the fit of the stochastic model by comparing the observed dataset to replicated datasets from the posterior predictive distribution. A visual comparison is shown in Figure 14. The model assumes that a single underlying dimension is sufficient to explain the choices made by the subjects. Hence, it is a model violation if subjects base their choices on a combination of two or more underlying characteristics. Previous analyses with other models (Van Mechelen and Van Damme, 1994, Leenen, Van Mechelen, and De Boeck, 1999) of this data set have shown that most subjects consider both the price and the available comfort in choosing a room and that both aspects are combined conjunctively (i.e., a subject selects a room if and only if it meets the subject's requirements with respect to both aspects). For the present data set, we will formally check the assumption of unidimensionality using posterior predictive checks.

Therefore, we first introduce the notation $z_2 = (q^{(1)}, a^{(1)}, b^{(1)}, q^{(2)}, a^{(2)}, b^{(2)})$ for a two-dimensional conjunctive parallelogram model (Leenen and Van Mechelen, 1999). The parameter vector z_2 contains two intervals $[a_i^{(1)}, b_i^{(1)}]$ and $[a_i^{(2)}, b_i^{(2)}]$ for each subject i and two positions $q_j^{(1)}$ and $q_j^{(2)}$ for each room j. The predicted data under this model are then defined by,

$$\hat{y}_{ij}(z_2) = \begin{cases} 1 & \text{if } a_i^{(1)} \leq q_j^{(1)} \leq b_i^{(1)} \text{ and } a_i^{(2)} \leq q_j^{(2)} \leq b_i^{(2)} \\ 0 & \text{otherwise.} \end{cases}$$

Let us now define the test statistic $T(y)$ as follows:

$$T(y) = D(y, \hat{y}(z_1(y))) - D(y, \hat{y}(z_2(y))),$$

where $z_h(y)$ $(h = 1, 2)$ is the best-fitting h-dimensional conjunctive parallelogram model found by applying Leenen and Van Mechelen's (1999) algorithm to y. As two- and one-dimensional models are nested, the proposed test statistic is formally similar to a likelihood ratio test (see also Rubin and Stern, 1994). If a data set y arises from a two-dimensional model, $T(y)$ is expected to be larger compared to a data set for which a one-dimensional model holds.

We computed the realized value $T(y)$ and compared it to the replicated values $T(y^{\text{rep}\,l})$ for the simulation draws $l = 1, \ldots, 4500$. Figure 12 shows the distribution of the replicated values and locates the realized value within that distribution: 22% of the replicated values exceed the realized value, showing no statistical evidence for the need for a two-dimensional model to underlie the current room data.

6 Discussion

Deterministic models are commonly used in many areas and so it is important to understand them from a statistical perspective. We do so here for models for binary data by considering minimal stochastic extensions of existing deterministic models. This involves three tasks: (1) setting up a probability model based on deterministic predictions and random errors, (2) creating a family of stochastic simulation algorithms that generalize existing error-minimization algorithms, and (3) setting up a framework for checking the fit of the new model by comparing observed data to replications under the model. Regarding (1), we set up two forms of a probability model from a Bayesian perspective, first, assuming equal error probabilities for all predictions and, second, allowing the error probability to depend on the predicted response. For (2), a Metropolis-Hastings algorithm can be constructed that elaborates on the steps of any existing deterministic optimization algorithm. Finally, the model checking steps (3) can be performed using posterior predictive checks.

Our general approach is illustrated for two models: a Guttman-scaling model fit to two sets of simulated item-response data, and a parallelogram model fit to real datasets from an archaeological study and from a study of individual choices.

From our simulated-data example, we learned that the stochastic model can recover information in the data that is not captured well by the best-fitting deterministic solution. In particular, the deterministic solution tends to understate the error rate and is poor at distinguishing the error rates from predicted 0's and 1's, even when these rates differ quite a bit in the data. By comparison, the stochastic models can recover these rates well provided the data originate from a model that is close enough to the fitted models. We were further particularly encouraged by the ability of the stochastic models to reveal systematic model violations through posterior predictive checks, which worked well both when the fitted model was approximately correct and when it was not.

From our real-data examples, we learned once again that, compared with the deterministic model, the stochastic model allows for a larger error rate, and a larger difference between errors in predicted 0's and 1's. This implies a more realistic understanding of how well the model really works (which would be relevant for predictive use of the model) as well as in what contexts it works better or worse. We also found that posterior predictive checks could find a model violations that would be hard to discover using only the deterministic fit.

We hope that our approach will be useful in other statistical problems. Leenen, Van Mechelen, and Gelman (2000) have applied this method for Boolean classification rules for binary data. Also, the problem of setting up a stochastic version of a deterministic model comes up in continuous-data problems (for example, Section 3.1 of Gelman and Bois, 1997, or, in a familiar setting, the idea of Gaussian errors and linear regression as a generalization of deterministic least squares algorithms).

References

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, **19**, 269–290.

Arabie, P., and Hubert, L. (1992). Combinatorial data analysis. *Annual Review of Psychology*, **43**, 169–203.

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

Degeest, R., Ottenburgs, R., Viaene, W., Kucha, H., Laduron, D., Bocquet, A., and Waelkens, M. (1997). Characterization of the common wares manufactured in Roman Sagalassos. An overview. In *Sagalassos IV. Report on the survey and excavation campaigns of 1994 and 1995. Acta Archaeologica Lovaniensia Monographiae 9*, ed. M. Waelkens and J. Poblome, 519–531. Leuven: University Press.

Fischer, G. H., and Molenaar, I. W., eds. (1995). *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer.

Fulkerson, D. R., and Gross, O. A. (1965). Incidence matrices and interval graphs. *Pacific Journal of Mathematics*, **15**, 835–855.

Gelman, A., and Bois, F. Y. (1997). How can statistical theory help with statistical practice? Example of a Bayesian analysis in toxicokinetics. In *Good Statistical Practice. Proceedings of the 12th International Workshop on Statistical Modelling*, ed. C. E. Minder and H. Friedl, 61–70. Wien: Austrian Statistical Society.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

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

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

Ghosh, S. P. (1972). File organization: the consecutive retrieval property. *Communication of the Association for Computing Machinery*, **9**, 802–808.

Gilks, W. R., Richardson, S., and Spiegelhalter, D., eds. (1996). *Practical Markov Chain Monte Carlo.* London: Chapman and Hall.

Goldmann, K. (1971). Some archaeological criteria for chronological seriation. In *Mathematics in the Archaeological and Historical Sciences: Proceedings of the Anglo-romanian Conference*, ed. F. R. Hodson, D. G. Kendall, and P. Tăutu, 202–208. Edinburgh: University Press.

Goodman, L. A. (1959). Some statistical methods for scalogram analysis. *Psychometrika*, **24**, 29–43.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association*, **70**, 755–768.

Guttman, L. (1950). The basis for scalogram analysis. In *Measurement and Prediction*, ed. S. A. Stouffer et al. Princeton: Princeton University Press.

Hoffman, R. J. (1979). On testing a Guttman scale for significance. *Educational and Psychological Measurement*, **39**, 297–301.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika*, **55**, 641–656.

Kendall, D. G. (1969). Incidence matrices, interval graphs and seriation in archaeology. *Pacific Journal of Mathematics*, **28**, 565–570.

Leenen, I., and Van Mechelen, I. (1999). A conjunctive parallelogram model for pick any/n data. Manuscript submitted for publication.

Leenen, I., Van Mechelen, I., and De Boeck P. (1999). Models for ordinal hierarchical classes analysis. Manuscript submitted for publication.

Leenen, I., Van Mechelen, I., and Gelman, A. (2000). Bayesian probabilistic extensions of a deterministic classification model. *Computational Statistics*, to appear.

Mitchell, S., and Waelkens, M. (1988). Cremna and Sagalassos 1987. *Anatolian Studies*, **38**, 53–64.

Poblome, J. (1996). Production and distribution of Sagalassos red slip ware. A dialogue with the Roman economy. In *Hellenistische und kaiserzeitliche Keramik des östlichen Mittelmeergebietes*, ed. M. Herfort-Koch, U. Mandel and U. Schädler, 75–103. Frankfurt am Main: Imbescheidt KG.

Poblome, J. (1999). *Sagalassos Red Slip Ware. Typology and Chronology. Studies in Eastern Mediterranean Archaeology 2.* Turnhout: Brepols.

Poblome, J., Viaene, W., Kucha, H., Waelkens, M., Laduron, D., and Depuydt, F. (1997). The clay raw materials of Sagalassos red slip ware. A chronological evaluation. In *Sagalassos IV. Report on the survey and excavation campaigns of 1994 and 1995. Acta Archaeologica Lovaniensia Monographiae 9*, ed. M. Waelkens and J. Poblome, 507–518. Leuven: University Press.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika*, **35**, 73–78.

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

Rubin, D. B., and Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In *Latent variable analysis: Applications for developmental research*, ed. A. von Eye and C. Clogg, 420–438. Thousand Oaks, Calif.: Sage.

Torgerson, W. S. (1962). *Theory and Methods of Scaling.* New York: Wiley.

Van Mechelen, I., and Van Damme, G. (1994). A latent criteria model for choice data. *Acta Psychologica*, **87**, 85–94.

Waelkens, M., and the Sagalassos team (1997). Interdisciplinarity in classical archaeology. A case study: the Sagalassos Archaeological Research Project (southwest Turkey). In *Sagalassos IV. Report on the survey and excavation campaigns of 1994 and 1995. Acta Archaeologica Lovaniensia Monographiae 9*, ed. M. Waelkens and J. Poblome, 225–252. Leuven: University Press.

Model	Error rates for data simulated from approx Guttman model			Error rates for data simulated from Rasch model		
	for pred 0's	for pred 1's	average	for pred 0's	for pred 1's	average
Determi-nistic	.214	.085	.145	.074	.107	.091
	.210	.084	.142	.090	.100	.095
	.200	.095	.142	.079	.112	.096
	.197	.101	.144	.077	.118	.099
	.211	.080	.141	.079	.110	.095
Truth	.250	.075	.162	.175	.168	.171

Table 1: Error rates for responses predicted to be 0, responses predicted to be 1, and all responses, for the two simulated-data examples of Section 4. The deterministic Guttman model was fit to both data sets shown in Figure 1. For each, error rates are shown for the true parameters and for the results of 5 different runs of the optimization algorithm. The deterministic algorithm mildly underestimates the error rate for the approximate Guttman data and drastically underestimates the error rate for the Rasch data.

Model	Data simulated from approx Guttman model				Data simulated from Rasch model			
	Error rates for		Error probabilities for		Error rates for		Error probabilities for	
	pred 0's	pred 1's	pred 0's	pred 1's	pred 0's	pred 1's	pred 0's	pred 1's
One error probability	.228	.098	.134		.099	.123	.115	
	.226	.098	.151		.079	.140	.126	
	.225	.102	.169		.114	.114	.115	
	.213	.111	.161		.103	.119	.105	
	.242	.094	.153		.091	.138	.113	
Two error probabilities	.281	.041	.298	.036	.085	.143	.083	.145
	.254	.069	.252	.056	.045	.167	.042	.167
	.260	.052	.217	.044	.079	.138	.095	.099
	.264	.067	.259	.053	.115	.110	.159	.107
	.265	.058	.270	.063	.102	.129	.111	.117
Truth	.250	.075	.250	.075	.175	.168	.171	.171

Table 2: Error rates and error probabilities for responses predicted to be 0 and responses predicted to be 1, for the two stochastic Guttman models fit to the two data sets shown in Figure 1. For each model and each data set, error rates are shown for the true parameters and for the results of 5 independent draws from the posterior distribution. For the one-error-probability model, the two error *rates* can differ substantially even though the two error *probabilities* are constrained to be equal. The stochastic models estimate the overall error rates well for the approximate Guttman data, but they substantially underestimate the error rate for the Rasch model (although not performing as badly as the deterministic model shown in Table 1).

1 0 0 **1** 0 **1** 0 0 **1** **1** **1** 0 0 0 0 0 **1** 0 0 0	0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 **1** **1** 0 0 0 **1** 0 **1** 0 0 0 0 0 0 0 0 0 0 0	**0** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 **1** 0 0 **1** 0 0 0 0 **1** 0 0 0 0 0 0 0 **1** 0	1 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 **1** 0 **1** **1**	**0** 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 **1** 0 0 0 **1** 0 0 0 **1** **1** 0 0 0	1 **0** 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 **1** **1** 0 **1** 0 0 0 0 0 0 0 0 0 **1** **1** **1** 0	**0** **0** **0** **1** 0 **1** **1** 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 0 0 0 0 **1** **1** 0 0 0 0 0 0 0 0 0 **1** 0	**0** **0** **0** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 **1** 0 0 0 0 0 **1** 0 **1** 0 0 **1** **1** **1** 0 **1**	1 1 1 **0** 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0
0 1 1 1 0 **1** 0 **1** 0 0 0 0 0 0 **1** **1** 0 0 **1** 0	1 **0** 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 **0** 1 1 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0	1 **0** **0** 1 1 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 **1** 0 0 0 **1** 0 0 0 0 0 0 0 0 **1** **1**	1 1 1 1 **0** **1** 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0
1 1 1 1 1 1 **1** 0 **1** 0 **1** **1** 0 0 0 0 **1** 0 0 0	1 **0** 1 1 1 **0** **1** 0 **1** 0 0 0 0 0 **1** 0 0 0 0 0
1 1 1 1 1 1 0 **1** **1** 0 0 0 **1** **1** 0 0 0 0 **1** 0	1 1 1 1 1 1 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 **0** 0 0 0 0 0 0 0 0 0 **1** 0 0 0	1 1 1 1 1 1 1 **1** 0 **1** 0 0 0 **1** 0 0 0 0 0 0
1 1 1 1 1 1 1 0 0 0 **1** 0 **1** 0 0 0 0 0 **1** 0	1 1 1 1 1 **0** 1 0 0 **1** 0 0 **1** 0 0 0 0 0 0 0
1 1 1 1 **0** 1 1 1 0 0 0 **1** 0 0 0 0 0 0 **1** **1**	1 1 **0** 1 **0** 1 1 1 **1** **1** 0 **1** 0 0 **1** 0 0 0 0 0
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 **1** 0	1 1 1 1 1 1 **0** 1 **1** 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 0 0 **1** 0 0 0 **1** 0 0 0 0	1 1 1 1 1 1 1 1 1 **1** **1** **1** 0 **1** 0 **1** **1** 0 0 0
1 1 1 1 1 1 1 1 1 0 0 0 0 **1** 0 0 0 **1** 0 0	1 1 1 **0** 1 1 **0** 1 **0** **1** **1** 0 **1** **1** **1** 0 0 0 0 0
1 1 **0** 1 1 1 1 1 1 1 **1** 0 0 0 **1** 0 0 0 0 0	1 **0** 1 1 1 **0** 1 1 1 **0** 0 0 0 0 0 **1** 0 0 0 0
1 **0** 1 1 1 1 1 1 1 1 **1** 0 **1** 0 0 **1** 0 0 0 0	1 1 1 1 1 1 1 **0** 1 1 **1** **1** 0 **1** 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 0 **1** **1** **1** 0 0 **1** 0 0	1 1 1 1 1 1 1 1 **0** **0** **0** **1** **1** **1** 0 0 0 0 0 0
0 1 1 1 1 1 1 1 1 **0** 1 **1** 0 0 0 **1** **1** **1** **1** 0	1 1 1 1 1 1 1 **0** 1 **0** **0** **1** 0 0 **1** 0 0 0 0 0
1 1 **0** 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 **1** 0	1 1 1 1 **0** 1 1 **0** **0** 1 1 1 **1** 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 **0** 0 0 **1** 0 0 **1** 0 **1**	1 1 **0** 1 1 1 1 1 1 1 1 **0** **1** 0 **1** 0 **1** **1** 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 0 **1** **1** 0 0 0 0	1 1 1 1 1 **0** 1 1 1 1 1 **0** **0** 0 **1** 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 **1**	1 1 1 1 1 1 1 1 1 **0** 1 1 1 **1** 0 0 **1** **1** 0 0
1 **0** 1 1 1 1 1 1 1 1 1 **0** 1 1 0 **1** 0 0 0 0	1 1 1 1 1 **0** 1 1 1 1 1 1 1 **0** **1** 0 **1** 0 0 0
0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 **1** 0 0	1 1 1 1 1 1 1 1 1 1 1 1 1 **0** **1** 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 0 **1** 0 0 0	1 1 1 1 **0** 1 1 1 1 **0** 1 1 1 **0** 1 0 0 **1** 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 **1** 0 0 0 0	**0** 1 1 1 1 1 1 1 1 **0** 1 1 1 1 **0** 0 **1** 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **1** 0 0 0	1 1 1 1 1 1 1 **0** 1 1 1 1 1 **0** 1 1 **1** **1** **1** 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 0 0 **1** 0	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 0 0 **1** **1**
1 1 1 1 1 1 1 1 1 1 **0** **0** 1 1 **0** 1 1 0 0 0	1 1 1 1 1 1 1 1 1 **0** 1 1 1 **0** 1 **0** **0** 0 **1** 0
1 1 1 **0** 1 1 1 1 **0** **0** 1 1 1 1 1 1 1 0 0 **1**	1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 1 **0** **1** **1** **1**
1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 **0** 1 0 0	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** **0** **0** 0 **1**
1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** **0** 1 1 0 0	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **1** 0
1 1 **0** 1 1 1 1 1 1 1 **0** 1 1 1 1 1 **0** 1 1 0	1 1 1 1 1 1 1 1 1 1 1 **0** 1 **0** 1 1 1 1 1 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 1 **1**	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** **1**
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 **0** 1 1 1 1

Figure 1: Simulated responses of 40 individuals to 20 items in the simulated-data examples of Section 4. The data on the left are simulated from a model approximating the Guttman model with errors; the data on the right are simulated from a Rasch model. We fit the Guttman model to both data sets. For each display, errors from the prediction given the true parameter values are in bold.

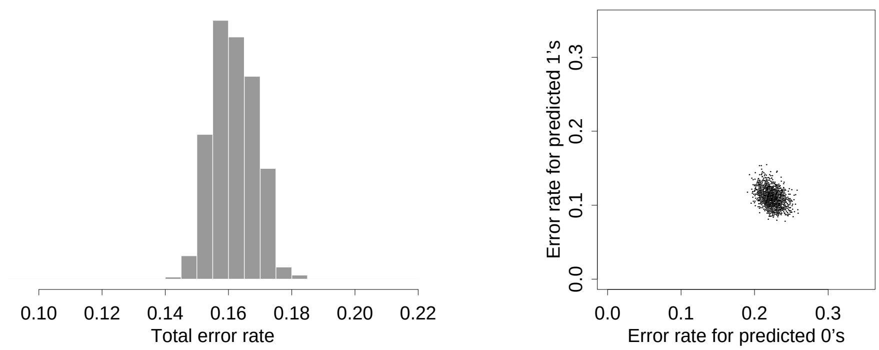

Figure 2: Posterior distribution of error rates from fitting the one-error-probability Guttman model to simulated approximate Guttman data: (a) histogram of 1000 posterior simulation draws of the total error rate $D(y, \hat{y}^z)/n$; (b) scatterplot of the conditional error rates $D_0(y, \hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$. Compare to a true total error rate of 0.162 and true conditional error rates of 0.250 and 0.075.

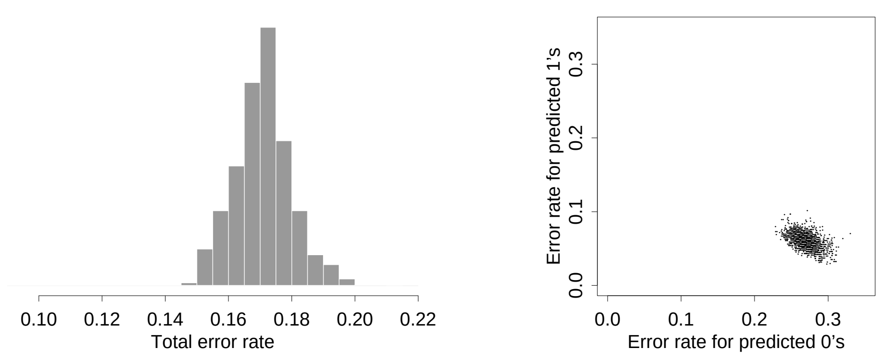

Figure 3: Posterior distribution of error rates from fitting the two-error-probability Guttman model to simulated approximate Guttman data: (a) histogram of 1000 posterior simulation draws of the total error rate $D(y, \hat{y}(z))/n$; (b) scatterplot of the conditional error rates $D_0(y, \hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$. Compare to a true total error rate of 0.162 and true conditional error rates of 0.250 and 0.075.

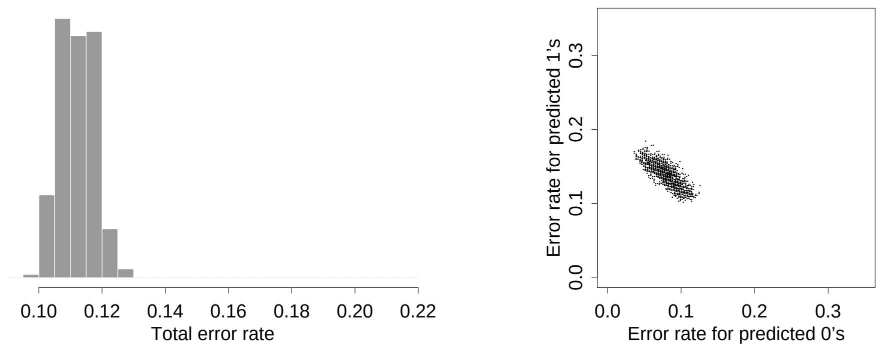

Figure 4: Posterior distribution of error rates from fitting the two-error-probability Guttman model to simulated Rasch data: (a) histogram of 1000 posterior simulation draws of the total error rate $D(y, \hat{y}(z))/n$; (b) scatterplot of the conditional error rates $D_0(y, \hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$. Compare to true total and conditional error rates of 0.17.

	NoN2−4	L10N−16N	L9S−18S	EoN11−18	EoN4−8	L8N−9N	L5N−7N	FDump	L3N−4N	Lib	HFill	LW9−14C	WDT	SoN5	EDT	LE4−6	HFloor	DT4
1C110	1	1	1	1	1	0	0	0	0	0	**1**	0	0	0	0	0	**1**	0
1C130−3	1	1	1	1	1	1	1	1	1	1	0	0	0	**1**	0	0	**1**	0
1A120	1	1	1	1	1	1	**0**	1	**0**	1	1	0	0	0	0	0	0	0
1C200−2	1	1	1	1	1	1	1	1	1	1	**0**	1	1	0	0	0	0	0
1A110−1	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	0	0
1A130	1	1	1	1	1	1	1	1	**0**	1	1	1	**0**	**0**	1	1	1	0
1B191	1	1	1	1	1	1	1	1	1	1	**0**	1	1	**0**	1	1	1	0
1C100	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1C120−3	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	1	1	0
1A100	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1
1B150	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	1	1
1B160−3	1	1	1	1	1	**0**	1	1	1	1	1	1	1	1	1	**0**	1	1
1B170	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	1	1
1B190	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1C170−1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	**0**	1	1	1
1F150	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1A160−1	0	1	1	1	1	1	1	**0**	1	1	0	0	0	0	0	0	**1**	0
1A150	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	**1**	**1**
1B260	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	1	0
1C190−1	0	0	1	1	1	1	1	1	**0**	1	1	0	0	0	**1**	0	0	0
1B251	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
1C180−3	0	0	0	1	1	**0**	1	**0**	1	1	1	1	1	1	1	1	1	1
1B220	0	0	**1**	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
1B100	0	0	0	0	0	0	1	1	1	1	1	1	**0**	1	1	1	0	0
1C160−1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
1C140	0	0	**1**	0	**1**	0	0	1	1	1	1	1	1	1	1	1	1	1
1A140−3	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
1B130	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
1F140	0	0	0	0	**1**	0	0	0	1	1	1	1	1	1	1	1	1	1
1F160	0	0	0	0	**1**	0	0	0	1	1	1	1	1	1	1	1	1	1
1F180	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	**1**
1B200	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0
1B210	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1B221	0	**1**	0	0	0	0	0	**1**	0	1	1	1	1	1	1	1	1	1
1B230−3	0	0	0	**1**	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1F170	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
1B180	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	**0**	1	0

Figure 5: Artifact × site data used in the example of Section 5.3. The artifacts i are ordered according to their leftmost interval parameters a_i under the best-fit deterministic model. Prediction errors are in bold.

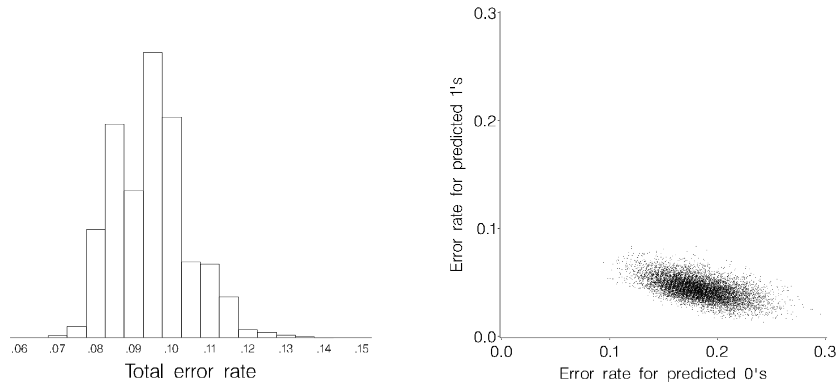

Figure 6: Posterior distribution of error rates from fitting the two-error-probability parallelogram model to the artifact $\times$ site data: (a) histogram of 10,000 posterior simulation draws of the total error rate $D(y, \hat{y}(z))/n$; (b) scatterplot of the conditional error rates $D_0(y, \hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y, \hat{y}(z))/n_1(\hat{y}(z))$.

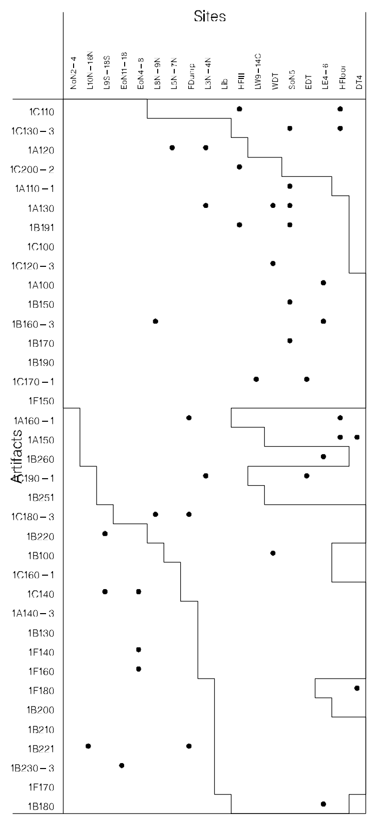

Figure 7: Graphical representation of the best fitting deterministic model for the artifact $\times$ site data. The lines indicate the boundaries of the artifacts' intervals. Sites are ordered according by their position and artifacts are ordered according to their left interval boundary. The dots represent mispredictions.

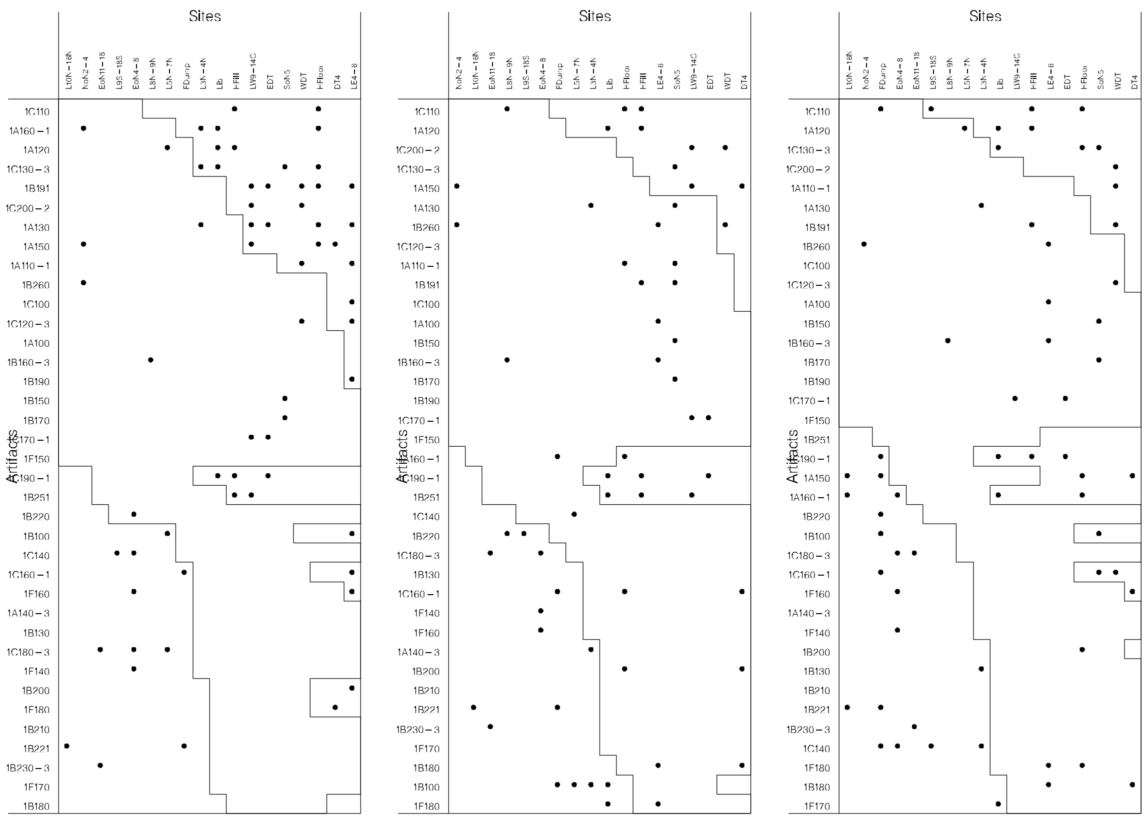

Figure 8: Graphical representation of three randomly selected draws from the posterior distribution for the artifact × site data. The lines indicate the boundaries of the artifacts' intervals. Sites are ordered according by their position and artifacts are ordered according to their left interval boundary.

1	1	1	1	1	1	1	1	**0**	1	1	1	1	1	1	0	0	0	0	0	0	**1**	0	0	0
1	**0**	1	1	**0**	1	1	1	**0**	1	1	1	1	**0**	1	1	1	0	0	0	**1**	0	0	0	0
0	1	1	1	**0**	1	1	1	1	1	1	0	0	**1**	0	**1**	0	0	0	0	0	**1**	0	0	0
0	1	1	1	1	**0**	1	**0**	1	1	1	**0**	1	1	1	0	0	0	0	**1**	**1**	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	1	**0**	1	1	0	0	0	0	**1**	**1**	0	0	**1**	**1**
0	0	1	1	1	1	1	**0**	1	1	1	1	1	1	**0**	1	1	1	0	0	0	0	**1**	0	0
0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	**1**	0	0	**1**
0	0	0	1	1	**0**	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1	1	**0**	1	1	1	0	0	0	0	0	0	**1**	0	0	0
0	**1**	0	0	0	1	1	1	1	1	1	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	**1**	**1**	0	0	0	0	**1**	**1**
0	0	0	0	**1**	0	0	1	1	1	1	**0**	1	1	0	0	0	0	**1**	0	**1**	0	0	0	0
0	**1**	0	0	0	0	0	0	1	**0**	1	1	1	1	0	0	0	**1**	0	0	0	0	0	0	0
0	0	0	0	0	0	**1**	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	**1**	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	**0**	1	1	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	1	**0**	1	1	1	1	1	1	1	1	1	1	1	0	0	0
0	**1**	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
0	0	0	**1**	**1**	0	0	0	0	0	1	1	1	1	0	0	0	**1**	0	0	0	0	0	0	**1**
0	**1**	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	**1**	0	0	0	0	0	0
1	0	**1**	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	**1**	**1**	0	0
0	**1**	0	0	0	0	**1**	0	0	0	0	0	1	1	0	0	0	**1**	**1**	0	0	0	0	**1**	0
0	0	0	0	0	0	0	0	0	**1**	0	0	1	1	1	1	**0**	1	0	0	0	0	0	**1**	0
0	0	**1**	**1**	**1**	0	0	0	0	0	0	0	1	1	1	1	1	1	1	**0**	1	1	1	1	1
0	0	0	0	0	**1**	0	0	**1**	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Figure 9: Student $\times$ room data used in the example of Section 5.4. The students i are ordered according to their leftmost interval parameters a_i under the best-fit deterministic model. Prediction errors are in bold.

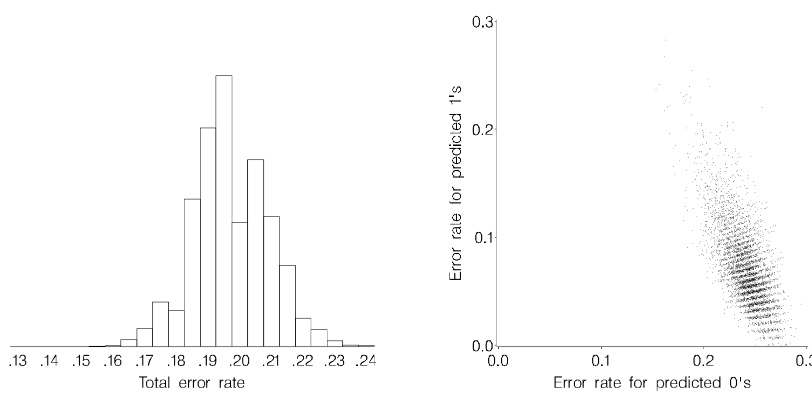

Figure 10: Posterior distribution of error rates from fitting the two-error-probability parallelogram model to student $\times$ room data: (a) histogram of 4500 posterior simulation draws of the total error rate $D(y,\hat{y}(z))/n$; (b) scatterplot of the conditional error rates $D_0(y,\hat{y}(z))/n_0(\hat{y}(z))$ and $D_1(y,\hat{y}(z))/n_1(\hat{y}(z))$.

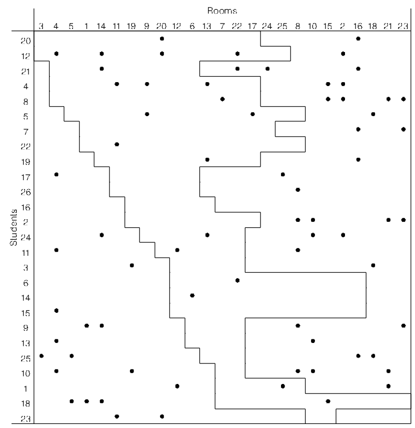

Figure 11: Graphical representation of the best fitting deterministic model for the student × room data. The lines indicate the boundaries of the students' intervals. Rooms are ordered according by their position and students are ordered according to their left interval boundary. The dots represent mispredictions.

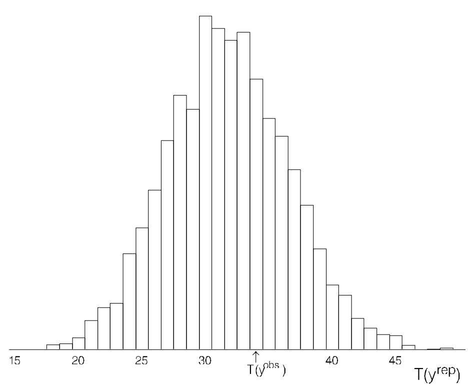

Figure 12: Posterior predictive check histogram of $T(y^{\mathrm{rep}})$ in the room example.

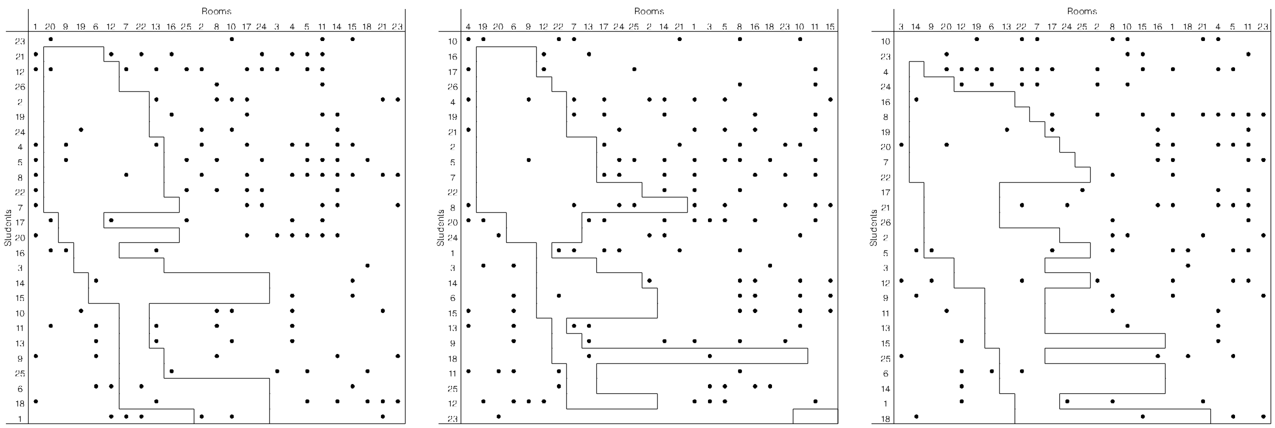

Figure 13: Graphical representation of three randomly selected draws from the posterior distribution for the student $\times$ room data. The lines indicate the boundaries of the students' intervals. Rooms are ordered according by their position and students are ordered according to their left interval boundary.

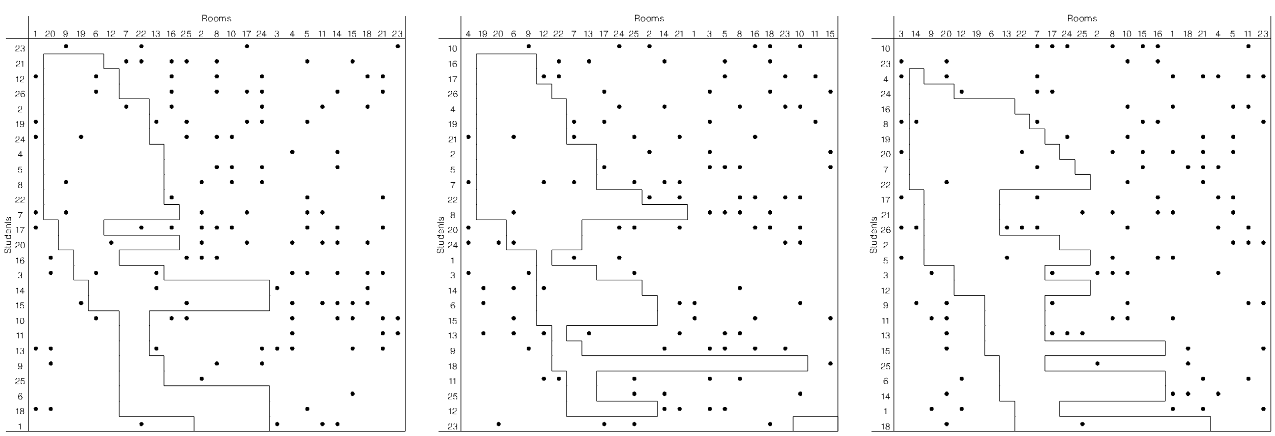

Figure 14: Graphical representation of three draws from the posterior predictive distribution corresponding to the three draws in Figure 13, respectively.