Rating Distributions and Bayesian Inference: Enhancing Cognitive Models of Spatial Language Use

Thomas Kluth
Language & Cognition Group
CITEC, Bielefeld University
Inspiration 1, 33619 Bielefeld
Germany
tkluth@cit-ec.uni-bielefeld.de

Holger Schultheis
Bremen Spatial Cognition Center
University of Bremen
Enrique-Schmidt-Str. 5, 28359 Bremen
Germany
schulth@uni-bremen.de

Abstract

We present two methods that improve the assessment of cognitive models. The first method is applicable to models computing average acceptability ratings. For these models, we propose an extension that simulates a full rating distribution (instead of average ratings) and allows generating individual ratings. Our second method enables Bayesian inference for models generating individual data. To this end, we propose to use the cross-match test (Rosenbaum, 2005) as a likelihood function. We exemplarily present both methods using cognitive models from the domain of spatial language use. For spatial language use, determining linguistic acceptability judgments of a spatial preposition for a depicted spatial relation is assumed to be a crucial process (Logan and Sadler, 1996). Existing models of this process compute an average acceptability rating. We extend the models and – based on existing data – show that the extended models allow extracting more information from the empirical data and yield more readily interpretable information about model successes and failures. Applying Bayesian inference, we find that model performance relies less on mechanisms of capturing geometrical aspects than on mapping the captured geometry to a rating interval.

1 Introduction

Acceptability judgments are an important measure throughout linguistic research (Sprouse, 2013). For instance, Alhama et al. (2015) recently proposed to use confidence ratings to assess models of artificial language learning. Likewise, in research on the evaluation of spatial language given visual displays, a common experimental paradigm is to ask how well a spatial term describes a depicted situation (e.g., Regier and Carlson, 2001; Logan and Sadler, 1996; Burigo et al., 2016; Hörberg, 2008). This paradigm results in individual acceptability judgments on Likert scales. These rating data are the main source for assessing computational models in the spatial language domain (e.g., Regier and Carlson, 2001; Coventry et al., 2005; Kluth and Schultheis, 2014). In other linguistic domains, similar empirical rating data are predicted by computational models (e.g., grammaticality judgments, Lau et al., 2017, or semantic plausibility judgments Padó et al., 2009; see also Chater and Manning, 2006).

Generally speaking, researchers consider a rating-model appropriate if it can closely account for empirical mean ratings for the given stimuli (averaged across subjects) – the closer the fit to the empirical mean data, the more appropriate the model. However, the use of mean ratings instead of full rating distributions misses the opportunity to use all available empirical information for model assessment. This is why we present a model extension that adds the simulation of a probability distribution over all ratings. We illustrate our extension by equipping spatial language models with full empirical rating distributions.

The second proposal of our paper (Bayesian inference) relies on the fact that our proposed model extension enables the generation of individual ratings by sampling from the simulated probability distribution. This opens up the possibility to apply Bayesian inference (e.g., to reason about the likely values of model parameters). Many cognitive models lack a likelihood function that specifies how likely the empirical data are given a specific parameter set. This prevents the use of Bayesian inference. In this contribution, we propose the

Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 47–55
Melbourne, Australia, July 19, 2018.

cross-match test developed by (Rosenbaum, 2005) as a means for computing the likelihood for cognitive models that are able to generate individual data.

Again, we use a spatial language model to exemplify the application of the cross-match method. The thus computed posterior distribution of the model's parameters has surprising implications for the interpretation of the model. Before we come to this, we start with presenting the example models, followed by our model extension to simulate rating distributions.

1.1 Exemplary Spatial Language Models

We introduce both our methods by exemplarily applying them to the AVS model (Regier and Carlson, 2001) and the recently proposed AVS-BB, rAVS, and rAVS-CoO models (Kluth et al., 2017, under revision). Given a depicted spatial layout and a spatial sentence ("The [located object] is above the [reference object]"), these cognitive models generate mean acceptability ratings, i.e., judgments how well the linguistic input describes the visual scene. All models can be interpreted as consisting of two components: One component that captures geometric aspects of the depicted spatial configuration and one component that maps the captured geometry to a rating interval (representing linguistic acceptability judgments).

The models process geometry by defining vectors on all points of one object of the spatial layout. These vectors point to the second object in the layout. In addition, each vector is weighted by a certain amount of attention defined by a spotlight-like distribution of attention. The overall direction of the vector *sum* is compared to a reference direction (e.g., canonical upright for the preposition *above*). This angular deviation is the outcome of the first model component (processing geometry).

The first model component is where the two model families (AVS & AVS-BB vs. rAVS & rAVS-CoO) differ: The AVS and the AVS-BB models assume a shift of attention from the reference object to the located object (the vectors point from the reference object to the located object). In contrast, the rAVS and rAVS-CoO models assume a reversed shift of attention from the located object to the reference object (hence their acronym: *reversed* AVS; the vectors point from the located object to the reference object). The difference within the model families (i.e., AVS vs. AVS-BB and rAVS vs. rAVS-CoO) will be introduced in Section 3.

The second model component is the same in all models: A linear function that maps the angular deviation from the first component to a rating interval. In Section 4.2.1 we introduce some details about the role of rAVS-CoO's parameters for the two model components. Applying our model extension and the second proposal of our paper (Bayesian inference), we present evidence that the second component of the models (mapping geometry to rating) seems to be more important than the first one (processing geometry).

2 Model Extension: Rating Distributions

As an illustrating example of our model extension, consider the empirical rating distribution displayed as bars in Fig. 1c. This distribution shows 34 acceptability ratings on a rating scale with $K = 9$ categories (from 1–9). These ratings come from an empirical study by Kluth et al. (under revision) in which they asked 34 participants to judge the acceptability of the German sentence "Der Punkt ist über dem Objekt" ("The dot is above the object"). Specifically, the distribution shown in Fig. 1c corresponds to empirical ratings for the left black dot above the asymmetrical object depicted in Fig. 1a.

Our method of simulating such a rating distribution is inspired by a common approach of analyzing ordinal data (i.e., discrete and ordered data) using generalized linear (regression) models (e.g., Liddell and Kruschke, 2018; Kruschke, 2015, chapter 23). Here, the cumulative probability of a latent Gaussian distribution between two thresholds is the probability of one specific rating k (see Fig. 1c).[1] Based on this, we propose the following steps to extend mean-rating-models with the ability of simulating full rating distributions:

1. Interpret the output of the model as the mean μ of a Gaussian distribution (see maximum of dashed curve in Fig. 1c or 1d).

2. Treat σ of the Gaussian distribution and $K - 1 - 2$ thresholds as additional model parameters (see width of dashed curve and vertical lines in Fig. 1c or 1d; K is the number of all outcomes; first and last thresholds have fixed values).

3. Define a discrete probability distribution over all K ratings like in an ordinal regression (i.e,

[1] For the first / last outcome it is the cumulative probability between negative / positive infinity and the first / last threshold.

cumulative probabilities of the Gaussian distribution between thresholds, see model outputs in Fig. 1c or 1d).

4. To generate an individual rating: Sample a rating from the discrete probability distribution defined in the previous step.

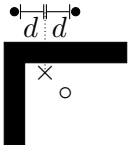
Note that the discrete probability distribution over all K ratings defined in step 3 is fully determined by the model parameters (i.e., it will not change unless you change any of the model parameters) while the individual rating generated in step 4 is subject to sampling noise.

To fit such an extended model to empirical data, we compute the Kullback-Leibler divergence from the model's discrete probability distribution (see model outputs in Fig. 1c or 1d) to the empirical rating distribution (relative frequencies of ratings, see bars in Fig. 1c or 1d) – for every dot-object pair that served as a stimulus. Then we minimize the mean Kullback-Leibler divergence (averaged over all stimuli). This procedure requires that individual empirical data are available.

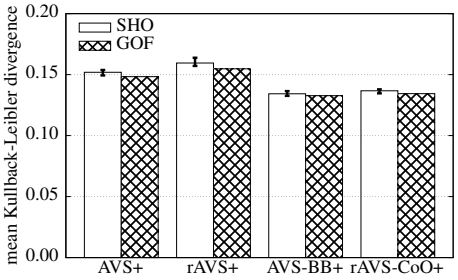
Note that this approach of comparing model outputs to empirical data still operates on the data from all study participants (but it uses more information as it does not operate only on a mean value). That is, instead of explicitly assessing the models on individual behavior, our fitting approach aims to capture the overall rating distribution. Given that with our model extension a model may also generate individual outcomes, it is in principle possible to explicitly model single individuals or groups of individuals with similar rating patterns. We leave this for future work and note that the work from Navarro et al. (2006) might prove valuable for this endeavor.

3 Results: Fitting Models to Rating Distributions

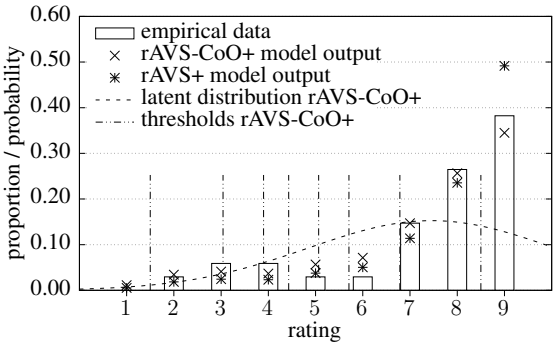
To exemplarily apply our proposed model extension, we extended the AVS model (Regier and Carlson, 2001) as well as the recently proposed AVS-BB, rAVS, and rAVS-CoO models (Kluth et al., 2017, under revision) and fitted them to empirical data from Kluth et al. (under revision, asymmetrical objects only). We denote the extended models with a trailing + (see labels in Fig. 1). The source code and all data are available under open licenses (GNU GPL and ODbL) from Kluth (2018).



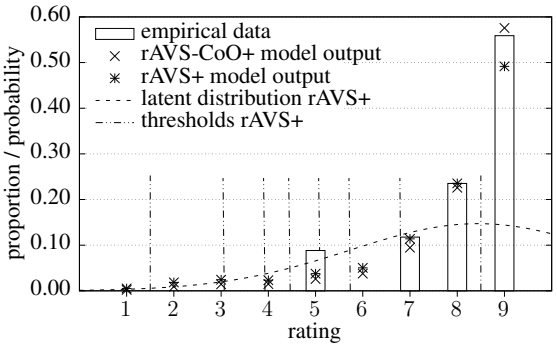
(a) Spatial configuration with two exemplary dot locations used in acceptability rating study by Kluth et al. (under revision). × = center-of-mass, ○ = center-of-object (of the asymmetrical object); d = same horizontal distance from × for both dots. Participants saw only one dot and the asymmetrical object (neither the centers nor the additional lines shown here).



(b) Goodness-of-fit (GOF) and simple hold-out (SHO) results for fitting extended models to whole empirical rating distribution from Kluth et al. (under revision, 4 asymmetrical objects × 28 dots × 2 prepositions = 224 data points). Error bars show 95% confidence intervals of SHO medians.



(c) Empirical "über" ("above") rating distribution and model probabilities (rAVS+ and rAVS-CoO+) for the **left dot** shown in Fig. 1a. Model probabilities were computed using the parameters from the best fit plotted in Fig. 1b. Participants never chose rating 1.



(d) Empirical "über" ("above") rating distribution and model probabilities (rAVS+ and rAVS-CoO+) for the **right dot** shown in Fig. 1a. Model probabilities were computed using the parameters from the best fit plotted in Fig. 1b. Participants never chose ratings 1-4 or 6.

Figure 1: Example experimental display, fits of extended models, and empirical rating distributions.

Given a depicted spatial configuration containing a geometric object and a single dot placed above / below the object (see Fig. 1a), we asked 34 German native speakers to rate the acceptability of the German sentences "Der Punkt ist über dem Objekt" and "Der Punkt ist unter dem Objekt" ("The dot is above / below the object") on a Likert scale from 1–9 (with lower ratings coding lower acceptability judgments). We placed 28 dots above and 28 dots below 4 asymmetrical objects (i.e., the whole data set consists of 224 data points; for the current work we did not consider data from additionally tested rectangular reference objects).

Fig. 1a shows two exemplary dot locations above one of the used asymmetrical objects. For these two dots, we expected participants to give equal "über" ("above") acceptability ratings (based on earlier research, e.g., Regier and Carlson, 2001). However, we found that participants rated the acceptability of the "über" ("above") sentence for the right dot in Fig. 1a higher than for the left dot (Kluth et al., under revision). This finding generalized reliably to different objects with similar dot placements suggesting that people possibly prefer the center-of-object (depicted as $\circ$ in Fig. 1a) over the center-of-mass (depicted as $\times$ in Fig. 1a) for their judgments. To account for this finding, Kluth et al. (under revision) proposed the model refinements AVS-BB and rAVS-CoO (AVS-bounding-box and rAVS-center-of-object), which both use the center-of-object instead of the center-of-mass (as AVS and rAVS do) for their computations.

Here, we use the two dot locations depicted in Fig. 1a to exemplarily present our approach of simulating rating distributions. To do so, we first extended all models with the ability to simulate rating distributions and then fitted all extended models to the 224 data points (by minimizing the mean Kullback-Leibler divergence as described above). These fits are plotted in Fig. 1b (as goodness-of-fit values alongside with the outcome of 101 simple hold-out iterations, a cross-validation measure to control for overfitting, Schultheis et al., 2013). In terms of relative model performances, these fits confirm the results of simpler fits using only averaged rating data reported in Kluth et al. (under revision): Both models that take the center-of-object into account (the AVS-BB+ and the rAVS-CoO+ models) fit the data more closely (lower mean Kullback-Leibler divergence) than the models that consider the center-of-mass (AVS+ and rAVS+).

More interesting for our current purpose are the plots in Figs. 1c and 1d. These plots each depict the empirical rating distributions for one of the two dots in Fig. 1a as bars: Fig. 1c shows the distribution for the left dot while Fig. 1d depicts the distribution for the right dot. The empirical distributions show that the left dot received considerably less "9" ratings and more "2–7" ratings compared to the right dot. On top of the empirical distributions, we plotted the probabilities of each rating as computed with the rAVS+ and the rAVS-CoO+ models. To compute these probabilities, we used the parameters found by fitting the models to the whole data set (cf. Fig. 1b). Despite being fit to a much larger data set, the two plots show that both models generally capture the qualitative trend of each of the two single empirical data points. Considering Fig. 1c and Fig. 1d suggests that the rAVS-CoO+ model better accounts for the data – confirming (and explaining, see Kluth et al., under revision) the better fit on the larger data set shown in Fig. 1b.

Fitting the models to rating distributions allows for a more fine-grained model assessment compared to model fits to averaged data. For example, the main source of the different performances of the rAVS+ and the rAVS-CoO+ models seems to be their ability to account for the frequency of the highest rating "9" (cf. Fig. 1c and Fig. 1d). Compare this with the situation where only averaged data is used: Here the only information are mean ratings (for the left dot 7.38, for the right dot 8.18) and fits of the models to these mean ratings. Using the same parameter settings as before, this yields for the left dot 0.1326 (rAVS fit, normalized root mean square error: nRMSE[2]) or 0.0093 (rAVS-CoO fit, nRMSE) and for the right dot 0.0333 (rAVS fit, nRMSE) or 0.1029 (rAVS-CoO fit, nRMSE). None of these numbers provides information about the models' properties as intuitive and informative as the fit of the extended models using full rating distributions. Moreover, our extension also enables the generation of individual data by sampling from the models' discrete rating distribution (see step 4 on page 3). This property can be used to analyze the models with Bayesian inference as we show next.

[2] $RMSE = \sqrt{\frac{1}{n}\sum_i^n (data_i - modelOutput_i)^2}$
$nRMSE = RMSE/(rating_{max} - rating_{min})$

4 Method: Bayesian Inference

The Bayesian framework is a fruitful and theoretically sound approach to reason with probability distributions over model parameters. However, this framework requires that the analyzed model can be interpreted in a probabilistic sense. As for many other cognitive models, this is not the case for any of the models discussed here (AVS, AVS-BB, rAVS, rAVS-CoO or their extended versions) because they lack a likelihood function that specifies how likely empirical data are given a model with a specific parameter set. We propose to use the cross-match test developed by Rosenbaum (2005) as the likelihood function of cognitive models that are able to generate individual data (e.g., the derivatives of the AVS+ model).

4.1 Cross-match Test

The cross-match test is a statistical test that computes the probability of whether multivariate responses of two differently treated subject groups come from the same distribution. In our case, the first group are empirical individual data and the second group are model-generated individual data (see top and bottom of Tab. 1), so the cross-match test becomes a measure of how likely it is that the model-generated data come from the same distribution as the empirical data. Given that we can only change the model-generated data (by using different parameter sets), this amounts to a likelihood function.

Internally, the cross-match test is based on grouping the multivariate responses (rows in Tab. 1) into pairs with minimal distances (Mahalanobis distances of ranks). The more of these pairs "cross-match" between the two groups, the more similar are the data of the two groups and hence the higher is the probability that the cross-match test computes (for more details see Rosenbaum, 2005).

4.2 Estimating the Posterior Distribution

To apply the cross-match test as a likelihood function of AVS+ derivatives, we propose the following procedure[3]:

1. For each stimulus, simulate as many ratings with the model as there were participants in the study by applying the procedure of generating individual ratings described in step 4 on page 3.
2. Compute the cross-match test comparing the empirical data with the model-generated data.
3. To account for sampling noise (see step 4 on page 3 in the generation of individual data) and provide reliable cross-match results for the same model parameters:
   (a) For every individual rating to be generated in step 1, sample s times and use the mean outcome as generated rating.
   (b) Use the following average of cross-match computations as likelihood value:
      i. Compute the mean number of cross-matches from c cross-match tests and store the probability for this number of cross-matches.
      ii. Repeat step i for b blocks and use the mean of these b probabilities as the likelihood value.

data type	left dot	right dot	...
empirical	7	8	...
empirical	9	9	...
...	...	...	...
model	8	9	...
model	5	8	...
...	...	...	...

Table 1: Example input for the cross-match test (Rosenbaum, 2005). Each row describes the response of one subject (empirical or model-generated), each column describes the response to a stimulus (e.g., the left or right dot from Fig. 1a).

Step 3 (b) basically repeats steps 1 and 2 $b \cdot c$ times. In our case, we found a sufficiently stable likelihood by applying step 3 with $s = 10$, $b = 20$, and $c = 4$ (standard error of averaged cross-match result < 0.05). Note that a too large value of s will generate model outputs that are too similar to each other and thus possibly reduces the number of cross-matches too much. The problem of an unstable likelihood value will reduce when more empirical individual data are available.

Having the likelihood function defined in this way, one can apply standard Markov Chain Monte

[3]Note that for clarity of presentation we stay in our exemplary domain: rating-models for spatial language. In principle, our approach is applicable to all models that are able to generate individual data points (not necessarily ratings).

Carlo (MCMC) techniques to estimate the posterior distribution. Specifically, we implemented a Metropolis-Hastings algorithm and improved its performance by adding the adaptation algorithm proposed by Garthwaite et al. (2016). For the cross-match test, we used the `R` package `crossmatch` (Heller et al., 2012) and re-implemented parts of it using the `C++` library `Armadillo` (Sanderson and Curtin, 2016). The `R` package `ggmcmc` (Fernández-i Marín, 2016) helped in visualizing and analyzing the MCMC samples. Again, all source code is available under the GNU GPL license from (Kluth, 2018).

4.2.1 Example rAVS-CoO+: Model Parameters & Prior Distributions

We exemplarily estimated the posterior distribution of the parameters of the rAVS-CoO+ model. The rAVS-CoO+ model has four free parameters (not considering the additional parameters of our ordinal model extension: σ and thresholds). The two parameters α and $highgain$ are part of the component that processes the geometry of the depicted spatial configuration (cf. Section 1.1). In particular α controls the extraction of an angular deviation from the spatial relation. This angular deviation is mapped to a linguistic rating with the second component of the model. Specifically, high angular deviation results in a low rating and low angular deviation results in a high rating. This is realized with a linear function that maps angular deviation to rating. The $intercept$ and $slope$ parameters are the parameters of this linear function.

Since this is the first study that investigates probability distributions over the model parameters of the rAVS-CoO+ model, we had no prior information available about the likely values of the model parameters. Accordingly, we used uniform distributions within the following parameter ranges as "uninformative" prior distributions:

$$\begin{aligned} \alpha \in\ & [0.001, 5];\ highgain \in [0, 10] \\ intercept \in\ & [0.7, 1.3];\ slope \in [-1/45, 0] \end{aligned}$$

5 Results: Bayesian Inference

We exemplarily estimated the posterior distribution of the parameters of the rAVS-CoO+ model for the same data set to which we fitted the model earlier (consisting of ratings for dots above / below asymmetrical objects, see Fig. 1b for model fits). We used 4 MCMC chains with 125,000 samples in each chain and checked the chains for convergence by monitoring the potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992). To obtain converging chains, we had to change the parameterization of the $slope$ parameter to measure "change per radian" instead of "change per degree". Furthermore, we kept the additional model parameters for the ordinal regression (σ of the latent Gaussian distribution and thresholds) constant on the values of the best rAVS-CoO+ fit to the whole data set, because we were primarily interested in the original model parameters. This parameter reduction improved the convergence of the MCMC chains while it did not affect the qualitative results. The results of the posterior estimation are plotted as density estimates of the marginal posterior distribution for each model parameter of the rAVS-CoO+ model in Fig. 2. The different colors code the different MCMC chains. The high overlap of the colors confirms the convergence of the chains.

At a first glance, the marginal posterior distributions are surprising as they lack clear maxima for any parameter in the considered ranges. In particular the α and the $highgain$ parameter seem to have little effect on the model output in terms of generating data similar to empirical data. On the other hand, the marginal posterior distributions suggest that the following regions in the parameter space should result in relatively poor model performance: $\alpha < 0.5$, $intercept > 1.0$, and $slope > -0.25$.

To double-check these regions, we picked two parameter sets and computed the model fits to the empirical data with these parameters (mean Kullback-Leibler divergence). The first parameter set lies in the presumably bad-performance region ($highgain = 5.0, \alpha = 0.2$, $intercept = 1.25$, $slope = -0.05$) while the second parameter set consists of parameter values from regions with high posterior density ($highgain = 5.0$, $\alpha = 3.0$, $intercept = 0.9$, $slope = -0.625$). Indeed, the presumably bad-performing parameter set fits the data worse than the other parameter set (mean Kullback-Leibler divergence: 0.484 vs. 0.266, respectively). This trend was confirmed with fits of the same parameter sets using mean ratings instead of rating distributions (nRMSE for worse parameters 0.301 vs. 0.145 for better parameters). These tests provide evidence that using the cross-match test as a likelihood function appropriately captures model performance.

After establishing the validity of the unexpected results, we discuss what we can learn from them.

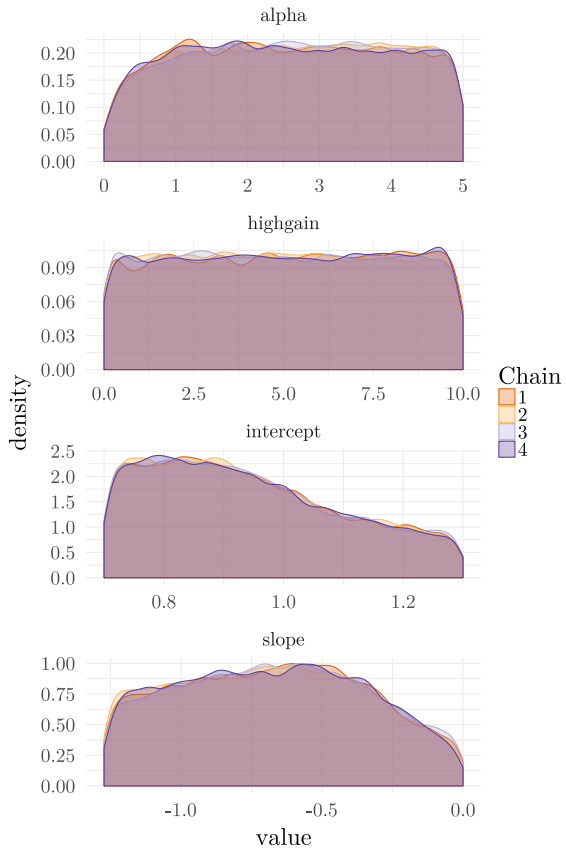


Figure 2: Marginal posterior distributions for the rAVS-CoO+ model given rating data from Kluth et al. (under revision, asymmetrical objects only) and "uninformative" prior distributions (uniform distributions).

Keep in mind that the following conclusions are only valid for the exemplary data set and model for which we computed the posterior estimation and may change with data highlighting different aspects of spatial language use.

Despite the great range of the parameter *highgain* its value does not affect the model performance. Accordingly, the parameter *highgain* seems to be irrelevant for the quality of the model output. Almost the same is true for the parameter α, although the marginal posterior distribution shows weak performance for values less than 0.5. The role of the parameter α in the rAVS-CoO+ model can be understood as an importance weight of two geometric features known to affect spatial language acceptability judgments: the proximal orientation and the center-of-object orientation (Regier and Carlson, 2001; Kluth et al., under revision). The closer α is to 0.0, the more important gets the proximal orientation and the less important gets the center-of-object orientation for the rAVS-CoO+ model. Thus, the marginal posterior distribution provides evidence that the center-of-object orientation is more important than the proximal orientation to account for this data set.

The *intercept* and *slope* parameters control the second model component (cf. Section 1.1): they are the parameters of a linear function contained in the rAVS-CoO+ model that maps angular deviation to rating (between 0 and 1). These two parameters have a greater influence on model performance than α and *highgain* (more diverse posterior profiles for *intercept* and *slope* compared to α and *highgain*, see Figure 2). That is, changing the values of the *intercept* or *slope* parameters affects the models' ability to fit empirical data more strongly than changing the values of α or *highgain*.

This is interesting, because one can interpret the rAVS-CoO+ model (and related models such as AVS+, AVS-BB+, rAVS+) as consisting of (i) a geometric component (capturing / formalizing the geometric properties of the involved objects and their spatial relation) and (ii) a mapping component (mapping the captured geometric aspects onto a rating range, see Section 1.1). Given that one of the prime research question motivating the development of these models concerns the influence of geometric properties (such as relative spatial location of the objects or asymmetrical objects) on spatial language use, most researchers focused on the geometric component of the models. Our results, however, suggest that the geometric component may be less important for model performance than commonly assumed – in particular, less important than the mapping component. That is, to unravel effects of geometry on spatial language use, it might be more insightful to re-consider the mapping of assumed intermediate geometric representations (e.g., angular deviations) to linguistic judgments instead of modeling the computation of these representations.

6 Discussion & Conclusion

Acceptability judgments are common in linguistic research (Sprouse, 2013). Many cognitive models of linguistic processes compute mean acceptability ratings. We propose a model extension that enables these models (i) to simulate a probability distribution over all possible ratings and (ii) to generate individual ratings. To fit simulated probability distributions to empirical rating distributions, we

propose to minimize the mean Kullback-Leibler divergence from the simulated to the empirical distributions. This model extension moves the model fits on a level that is closer to the actual empirical data (by using full rating distributions instead of mean ratings) while it avoids the problematic treatment of ordinal data as metric (Liddell and Kruschke, 2018). As future steps in this direction, we envision an analysis whether the additional model parameters can be mapped onto cognitive structures and mechanisms and subsequently the explicit modeling of (groups of) individuals (e.g, via Navarro et al., 2006).

Since many cognitive models lack a likelihood function, our additional contribution is to introduce the cross-match test (Rosenbaum, 2005) as a possible approximation of the likelihood function. This adds the possibility to apply full Bayesian inference for the parameters of all cognitive models that are able to generate individual data (e.g., mean-rating-models enhanced with our model extension).

In the related work of Approximate Bayesian Computation (ABC, for review see Turner and Van Zandt, 2012), researchers have developed sampling strategies to enable "likelihood-free inference". These techniques enable a modeler to use the Bayesian toolkit without explicitly defining a likelihood function. However, ABC sampling algorithms add additional overhead to the workflow of cognitive modelers, as they diverge from standard MCMC techniques used in Bayesian estimations. To overcome this overhead, we propose to use the cross-match test as an explicit likelihood function. We are currently evaluating our approach in comparison to existing ABC algorithms.

We exemplarily applied both our proposals using computational cognitive models of spatial language use like the AVS model (Regier and Carlson, 2001) and its derivatives (Kluth et al., 2017, under revision). Given a depicted spatial layout and a spatial preposition, these models compute mean acceptability ratings. We showed that simulating rating distributions allows a more fine-grained model assessment compared to model fits using mean ratings.

An example application of Bayesian inference revealed surprising insights: We estimated the posterior distribution of rAVS-CoO+'s parameters and found that the values of almost all parameters were less important for model performance than we thought. Future research in this direction will help to precisely identify and quantify the role of model parameters for the rAVS-CoO+ model (and the related models AVS+, AVS-BB+, and rAVS+). In addition, the Bayesian toolkit comprises several other methods for model inspection and model comparison.

Acknowledgments

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Raquel G. Alhama, Remko Scha, and Willem Zuidema. 2015. How should we evaluate models of segmentation in artificial language learning? In *Proceedings of the 13th International Conference on Cognitive Modeling*.

Michele Burigo, Kenny R. Coventry, Angelo Cangelosi, and Dermot Lynott. 2016. Spatial language and converseness. *Quarterly Journal of Experimental Psychology*, 69(12):2319–2337.

Nick Chater and Christopher D Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *Spatial Cognition IV. Reasoning, Action, Interaction*. Springer.

Paul H. Garthwaite, Yanan Fan, and Scott A. Sisson. 2016. Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics-Theory and Methods*, 45(17):5098–5111.

Andrew Gelman and Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Ruth Heller, Dylan Small, and Paul Rosenbaum. 2012. *crossmatch: The cross-match test*. R package version 1.3-1.

Thomas Hörberg. 2008. Influences of form and function on the acceptability of projective prepositions in Swedish. *Spatial Cognition & Computation*, 8(3):193–218.

Thomas Kluth. 2018. A `C++` implementation of cognitive models of spatial language understanding as well as pertinent empirical data and analyses. will soon be published under `https://pub.uni-bielefeld.de/person/54885831/data`.

Thomas Kluth, Michele Burigo, and Pia Knoeferle. 2017. Modeling the directionality of attention during spatial language comprehension. In Jaap van den Herik and Joaquim Filipe, editors, *Agents and Artificial Intelligence*, Lecture Notes in Computer Science. Springer International Publishing AG.

Thomas Kluth, Michele Burigo, Holger Schultheis, and Pia Knoeferle. under revision. Does direction matter? Linguistic asymmetries reflected in visual attention. *Cognition*.

Thomas Kluth and Holger Schultheis. 2014. Attentional distribution and spatial language. In Christian Freksa, Bernhard Nebel, Mary Hegarty, and Thomas Barkowsky, editors, *Spatial Cognition IX*, Lecture Notes in Computer Science. Springer.

John K. Kruschke. 2015. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*, 2nd edition. Academic Press.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Torrin M. Liddell and John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? Preprint, retrieved from osf.io/9h3et.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merill F. Garrett, editors, *Language and Space*, chapter 13. The MIT Press.

Xavier Fernández-i Marín. 2016. ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, 70(9):1–20.

Daniel J. Navarro, Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. 2006. Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Paul R. Rosenbaum. 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530.

Conrad Sanderson and Ryan Curtin. 2016. Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software*, 1:26.

Holger Schultheis, Ankit Singhaniya, and Devendra Singh Chaplot. 2013. Comparing model comparison methods. In *Proc. of the 35th Annual Conference of the Cognitive Science Society*, pages 1294 – 1299, Austin, TX. Cognitive Science Society.

Jon Sprouse. 2013. Acceptability judgments. In *Oxford Bibliographies*. Oxford University Press.

Brandon M. Turner and Trisha Van Zandt. 2012. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85.