Exploring the Representation of Linear Functions

Pablo León-Villagrá[1] Verena S. Klar[1] Adam N. Sanborn[2] Christopher G. Lucas[1]
[1] School of Informatics, University of Edinburgh, United Kingdom
[2]Department of Psychology, University of Warwick, United Kingdom

Abstract

Function learning research has highlighted the importance of human inductive biases that facilitate long-range extrapolations. However, most previous research is focused on aggregate errors or single-criterion extrapolations. Thus, little is known about the underlying psychological space in which continuous relationships are represented. We ask whether people can learn the distributional properties of new classes of relationships, using Markov Chain Monte Carlo with People, and find that (1) people are able to track not just the expected parameters of a linear function, but information about the variability of functions in a specific context and (2) in many cases these spaces over parameters exhibit multiple modes.

Keywords: generalization, function learning, representation

Inductive biases are at the heart of the human ability to generalize and extrapolate from sparse evidence. For instance, when we infer labels and properties of new objects or entities, we rely not just on our experience of past examples, but on our implicit and explicit expectations about the nature of categories. Similarly, when we learn relationships between quantities in *function learning*, inductive biases make it possible to distinguish between the boundless possible relationships behind a set of observations (Lucas et al., 2012, 2015).

In order to characterize a person's inductive biases, it can be useful to first focus on spaces of possible mental representations – sometimes called hypothesis spaces – and the kinds of inferences they support or preclude. As with categorization, there have been many proposals about the mental representations supporting function learning, including exemplar-based approaches (McDaniel & Busemeyer, 2005), rule-based approaches (Brehmer, 1974), and hybrids or generalizations of these (DeLosh et al., 1997; Lucas et al., 2015). These models are typically evaluated by comparing their predictions to averaged human judgments, either via direct correlations, error relative to the true underlying function, or qualitative features including multiple modes (Kalish et al., 2004), or monotonicity (Bott & Heit, 2004; Kalish, 2013).

While this line of research has shed light on function learning and the representations and inductive biases that make it possible, some fundamental questions remain. For example, while models that take a distributional approach to function learning have successfully explained human behavior, there is little direct evidence that people track distributional information – uncertainty or variability – when faced with function learning problems. This question has been unanswerable in previous work that relied on aggregated judgments or assumed that individual inductive biases are broadly similar (Kalish et al., 2007). Even the few studies that have focused on inference patterns (Kalish, 2013; Wilson et al., 2015; Schulz et al., 2017), including analyses of per-participant extrapolations (León-Villagrá et al., 2018), still neglected this question about the tacit beliefs behind participants' judgments. Only recently, experiments have started to explore the role of uncertainty in function learning. In Schulz et al. (2015) participants judged functions to be more predictable when they were smooth or when they exhibited low variance, much in accordance with the preferences of a probabilistic model. Similarly, Stojic et al. (2018) showed that participants' predictive accuracy in a function learning task correlated with their confidence ratings, again resembling the uncertainty estimated by a probabilistic model.

Here we expand on this work and attempt to directly characterize how people represent uncertainty when they learn functions.

Markov Chain Monte Carlo with People

To uncover the psychological space that participants learn when learning functions we apply Markov Chain Monte Carlo with People (MCMCP; Sanborn et al., 2010). Sanborn et al. showed that Markov Chain Monte Carlo can be used as an experimental method to elicit posterior distributions from people using a simple forced-choice task. Thus, MCMCP offers a method to explore the psychological representational space and has been successfully applied to elicit the representations of complex stimuli, such as facial affect categories (Martin et al., 2012). Previously, MCMCP has been used in a function learning setting[1] to examine if participants prefer compositional over non-compositional functions (Schulz et al., 2017). Since Schulz et al. were interested in preferences for types of functions (compositional vs. non-compositional), the samples presented consisted of discrete varieties of functions and did not explore the distribution of function parameters.

In contrast, in this work, we directly explore the distributional space of the parameters governing the realizations of linear functions. This allows us to uncover how learned functions are represented, without constraining the participant's choices to pre-specified sets of materials.

Adopting MCMCP also allows us to explore novel questions – do participants represent variability in the training relationships? Do they form a single, deterministic functional relationship or do they form posterior distributions over parameters, reflective of the variability in the training? This question about representation, in turn, can inform more general future questions about extrapolation – are typical extrapolation patterns maximum a posteriori judgments given a

[1]Function learning has been more extensively studied in a closely related paradigm, *iterated learning*. Iterated learning experiments can elicit participants' shared expectations and have revealed strong inductive biases for positive linear functions (Kalish et al., 2007).

learned distribution over parameters? Or do they correspond to samples from a range of probable parametrizations?

In this work we:

- Evaluate if MCMCP can be successfully adapted to a function learning paradigm.
- Contrast how functions are represented depending on the variability of the example sets provided.

Experiment

In this experiment, we examine how participants represent linear functions when presented with sets of training examples. We hypothesize that participants learn both the parameters generating the function, as well as the variability of the relationship, i.e. they will learn both how much slopes and intercepts vary, while also learning the specific modes of slopes and intercepts. Therefore, we expect participants to form posterior distributions over the training parameters, with the variance of that posterior reflective of the training.

We distinguish between training functions with positive and negative slopes, since previous research has highlighted strong inductive biases for these relationships. Similarly, while it has been shown that people are biased to extrapolate in a linear fashion, especially preferring linear functions where both stimulus and criterion are matched (DeLosh et al., 1997), extrapolations appear to be influenced by their proximity to the extrapolation boundaries. In areas of the extrapolation range that are closer to zero, participants seem to adjust the slope of their extrapolations towards this boundary (Brown & Lacroix, 2017; Kwantes & Neal, 2006). To test how different offsets and different degrees of steepness are represented we contrast steep and shallow linear functions. Finally, we expect that highly salient functional relationships, like positive functions for which target and criterion are matched, will be easier to learn and result in more peaked posterior distributions if the training exhibits low variability. For high variability training, and especially if the function is not favored as strongly (for instance a function with a shallow negative slope) we expect broader, less peaked posteriors. Finally, we hypothesize that especially in high variability conditions, some participants will not exhibit unimodal posterior distributions and consider several potential generating functions broadly consistent with the learned function.

Contrasting these functions resulted in a $2 \times 2 \times 2$ between-subjects design (direction of the function: positive or negative, steepness: shallow or steep, variability of the training data: low or high).

Participants

The study was self-certified in accordance with the School of Informatics Ethics Guidelines. We recruited 454 participants ($M_{age} = 33$, $SD_{age} = 8.63$, 91 female, 176 male, 1 other, 186 refused information on gender) on Amazon Mechanical Turk. Participants had to have more than 50 approved HITs and an approval rate of 95% or larger. They received \$1.33 for participation and took an average of 17 minutes ($M = 17.25, SD = 8.59$) to complete the experiment. Participants were randomly assigned to one of the 8 conditions.

Materials

The parameters generating the functions in the experimental conditions differed in the sign of the slopes, as well as in their steepness. In addition, parameters in the training set exhibited either low or high variance for intercepts and slopes. For the full set of experimental conditions, see Table 1.

Table 1: Parametrization for the generating linear functions.

Condition	β_0	SD_{β_0}	β_1	SD_{β_1}
$C_{.5,\text{low}}$	0.25	0.05	0.5	0.025
$C_{1.0,\text{low}}$	0	0.05	1	0.025
$C_{-.5,\text{low}}$	0.75	0.05	-0.5	0.025
$C_{-1.0,\text{low}}$	1	0.05	-1	0.025
$C_{.5,\text{high}}$	0.25	0.3	0.5	0.15
$C_{1.0,\text{high}}$	0	0.3	1	0.15
$C_{-.5,\text{high}}$	0.75	0.3	-0.5	0.15
$C_{-1.0,\text{high}}$	1	0.3	-1	0.15

To create the 25 training sets, corresponding to iid realizations of $\beta_0, \beta 1 \sim \mathcal{N}(\mu, \sigma)$, with μ and σ matching the experimental condition, we systematically sampled 10,000 pairs and selected the most normal and uncorrelated sets[2]. Then we generated the corresponding linear function for a range of 15 points for x in 0–1 for all sets. One of those 15 values was picked at random and constituted the interpolation target.

MCMCP Proposals were generated by two symmetric Gaussian distributions, to allow both for local, as well as far-off proposals, $\sigma_{\beta_0} \in [0.14, 0.98]$, $\sigma_{\beta_1} \in [0.21, 1.47]$. At each iteration these proposals had a probability of .8 and .2 to be selected. Proposals were further restricted to be in bounds $\beta_0 \in [-0.5, 1.5]$, $\beta_1 \in [-1.5, 1.5]$, and if less than 4 points of the function realization were visible on screen, the proposal was automatically rejected and a new proposal was resampled. Participants traversed three different, interleaved chains, since multiple chains allow a wider application of convergence diagnostics and reduce the impact of the particular starting state. The starting values for these chains were obtained by k-means clustering of pilot data ($n = 8$, one participant per condition). This resulted in the following starting values $\beta_0 = \{0.12, 0.1, 0.58\}, \beta_1 = \{0.92, -0.94, -0.28\}$, for chains 1 to 3.

Procedure

Participants were instructed that they would learn the relationship between two proteins, Zenopin and Mepradin. Participants were told that the concentration of Zenopin was related to Mepradin, but that the extent of that relationship var-

[2] All Shapiro-Wilk tests yielded $p > 0.99$, and all correlation coefficients were in the range $[-.01\ .01]$.

ied between humans. Participants were also instructed that they would be presented with examples of the relationship as observed in different people and that they would be asked to interpolate the relationship. They were then instructed that after the training phase they would be presented with pairs of proposed relationships, all observed for a new person, and would have to choose which of the two were more likely to resemble the learned relationship. After reading the set of instructions, the participants were tested on their comprehension. If participants did not respond correctly in the questionnaire they had to restart the instructions.

Training Phase In the training phase, participants were presented with 25 interpolation tasks, presented as scatter plots. In each task, they were instructed that the scatter plot depicted the relationship between the two protein concentrations for a new person. They then had to guess the concentration of the protein by selecting the height of the corresponding value on the plot (on the y-axis). Participants were shown the correct value as feedback for one second, and, if their choice deviated by more than ± 0.05 from the true value, had to readjust their selection.

Test Phase The test phase consisted of 240 forced-choice tasks, corresponding to 80 interleaved iterations of the three Markov chains. On each trial, participants were presented with two adjacent scatter plots, one corresponding to the current state of the chain and the other reflecting the proposed new state (in randomized order). Participants had to select the plot they believed most likely to depict the relationship in the training phase. After the test phase, participants completed a short survey, were debriefed, and compensated. See Figure 1 for a depiction of both training and test phase.

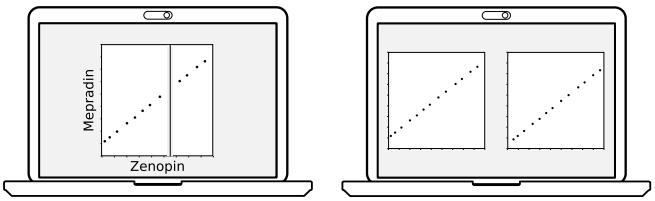

Figure 1: Participants had to complete a training and a test phase. In the training phase they were asked to interpolate the concentration of a fictitious protein for 25 different people (with feedback). In the test phase, they were presented with 240 forced-choice tasks, for which they had to choose the scatter plot that most resembled the relationship in the training phase. The choices were presented in random order and corresponded to a Markov chain, in which the participant implemented the acceptance function.

Results

We excluded participants from the analysis if their chains did not converge to the stationary distribution. Many criteria for convergence checks have been suggested in the literature, here we applied one of the most commonly used evaluations, $\hat{R}$ (Gelman et al., 2013; Vehtari et al., 2019). $\hat{R}$ estimates the ratio between within-chain variances and between-chain variance and thus provides a measure of how (self-)similar chains are. In general applications $\hat{R}$ should not exceed a value of 1.1. However, such a strict application of this diagnostic is not realistic in most MCMCP experiments, since human judgments might exhibit more correlated choices and the number of iterations in experiments is usually considerably lower than in standard statistical applications. Therefore, we incrementally calculated $\hat{R}$ values for chains for each participant and selected the lowest overall $\hat{R}$, with the additional constraint that the first 20 samples of the chain were always discarded and the resulting chains had to be at least 20 iterations long. We then used the maximum of the intercept and slope $\hat{R}$ values to apply exclusion criteria and determine burn-in.

Similar to Ramlee et al. (2017), we excluded participants who exhibited $\hat{R} \geq 2$. Furthermore, we excluded participants who required more than one correction in the interpolation task. Given that the interpolation function was deterministic, most participants did not require many corrections ($Mdn = 0, \mathrm{Q1} = 0, \mathrm{Q3} = 1, max = 44$).

In total, these methods led to the exclusion of 262 participants (convergence exclusions: 224, interpolation exclusions: 72). This high number of exclusions was to be expected given the correlated, bi-variate parameter space and previous results (Sanborn et al., 2010). For group sizes after exclusion, see Table 2. For an overview of how the forced-choice task results in the posterior distribution, see Figure 2.

Determining Burn-in

To determine how many trials were required on average for the Markov chains to converge, we used the iteration for which $\hat{R}$ was optimal for each participant. On average, chains required 33 iterations to reach optimal burn-in and the resulting optimal $\hat{R}$ values were well below 2, $M_{\hat{R}} = 1.4, SD = 0.2$. Conditions did not differ considerably in terms of the optimal iterations or the resulting $\hat{R}$ values. For the full list of per-condition burn-in values, see Table 2. For all subsequent analysis, we discarded all points of the chain before the per-participant burn-in.

Table 2: Participants in each condition before (N_{total}) and after exclusion (N). $M_{\text{burn-in}}$, $SD_{\text{burn-in}}$, as well as mean acceptance probabilities averaged over participants (M_{acc}, SD_{acc}).

Condition	N_{total}	N	$M_{\text{burn-in}}$	$SD_{\text{burn-in}}$	M_{acc}	SD_{acc}
$C_{.5,\text{low}}$	48	25	34.88	14.49	35	17
$C_{1.0,\text{low}}$	63	21	31.37	12.01	42	10
$C_{-.5,\text{low}}$	52	19	34.37	13.73	37	13
$C_{-1.0,\text{low}}$	64	22	29.59	11.59	38	15
$C_{.5,\text{high}}$	59	35	32.29	13.22	38	14
$C_{1.0,\text{high}}$	57	26	32.08	12.24	45	9
$C_{-.5,\text{high}}$	56	29	35.66	12.75	42	13
$C_{-1.0,\text{high}}$	55	15	29.40	10.67	36	12

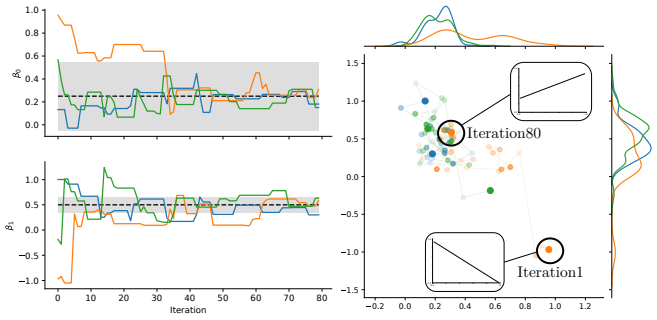

Figure 2: The 240 choices submitted by the participants corresponded to three Markov chains. By accepting or rejecting proposed parametrizations for the functions, participants traversed this representational space and eventually converge to a region reflecting the posterior over parameters. For this participant, the chains converge after 35 iterations for β_0 and after 15 iterations for β_1. The corresponding distribution after this burn-in period closely matches the true relationship learned in the training phase, both in terms of its mean and variance (dashed line and grey range).

Acceptance Probabilities

Acceptance rates for MCMC samples should range between 20–40% (Roberts, Gelman, & Gilks, 1997). Mean acceptance probability was in that range, $M = 39\%, SD = 13$, indicating that the proposals were wide enough to traverse the parameter space. Between conditions, the mean acceptance probabilities for participants varied, ranging from 35 to 45%, for all acceptance probabilities, see Table 2. For each condition, acceptance probabilities for each chain did not vary substantially and were similar to the general acceptance rates (not shown).

Posterior Distributions

Slopes differed significantly between positive- and negative-slope conditions, with participants trained on negative slopes preferring negative slopes, $M_{\beta_1} = -0.16, SD_{\beta_1} = 0.53$, and participants trained on positive slopes preferring positive slopes, $M_{\beta_1} = 0.19, SD_{\beta_1} = 0.45$, $t(165.33) = -4.74, p < .0001$[3].

For conditions with negative slopes in the training sets, steep and shallow conditions exhibited significantly different posterior slopes, with lower slopes for steep compared to shallow conditions, $M_{-.5} = -0.05, SD_{-.5} = 0.45, M_{-1.0} = -0.29, SD_{-1.0} = 0.59$, $t(65.58) = 2.08, p = .041$. For conditions with positive slopes in the training sets there was also a significant difference in posterior slopes. However, this difference was not in the predicted direction, as slopes in the shallow condition were on average larger than in the steep condition, $M_{.5} = 0.29, SD_{.5} = 0.4, M_{1.0} = 0.05, SD_{1.0} = 0.47$, $t(89.75) = -2.89, p = .005$. Posterior intercepts in conditions with negative training slopes did not differ significantly between steep and shallow conditions, $M_{-.5} = 0.52$, $SD_{-.5} = 0.21$, $M_{-1.0} = 0.6$, $SD_{-1.0} = 0.3$, $t(62.84) = 1.38$, $p = .174$, nor for conditions with positive training slopes, $M_{.5} = 0.35, SD_{.5} = 0.2, M_{1.0} = 0.5, SD_{1.0} = 0.25, t(88.71) = 3.31, p = .001$.

[3] All tests are unequal variance, two-sided t-tests.

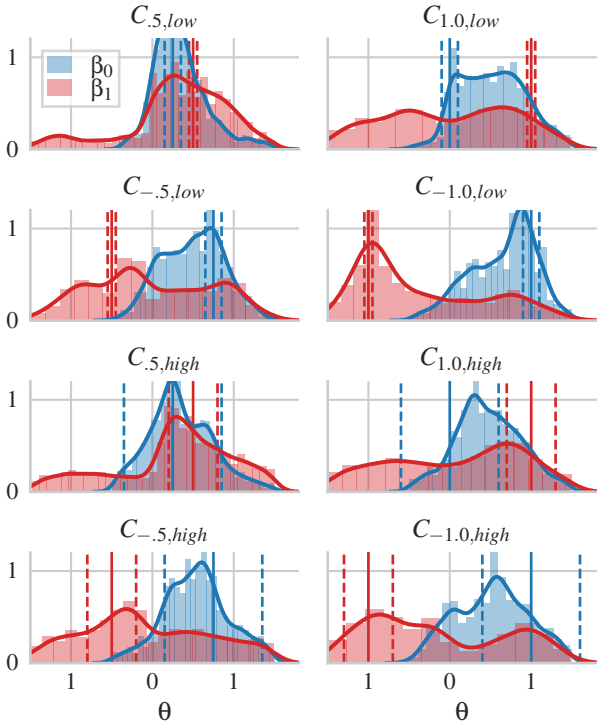

Figure 3: Posterior densities for intercepts and slopes and true values and standard deviation (dashed lines) in the experimental conditions. The posterior densities exhibited multiple modes, some centered in close proximity of the true parameters.

Equally, per-participant SD for slopes did not differ significantly between high and low variability conditions, $M_{low,\beta_1} = 0.49$, $SD_{low,\beta_1} = 0.26$, $M_{high,\beta_1} = 0.55$, $SD_{low,\beta_1} = 0.25$, $t(180.07) = -1.39$, $p = .166$. However, for intercepts, per-participant SD did differ significantly between high and low variability conditions, with high variance conditions resulting in higher SD, $M_{low,\beta_0} = 0.26, SD_{low,\beta_0} = 0.11, M_{high,\beta_1} = 0.31, SD_{low,\beta_1} = 0.11, t(182.48) = -2.46, p = .015$.

Visual inspection revealed that in all conditions posterior distributions were multimodal and heavily skewed, which complicated the analysis. In general, the posterior densities suggested that the modes of the posterior distributions were often close to the learned parameters, see Figure 3, for a selection of posterior distributions for one participant in each condition, see Figure 4.

Since the mean and standard deviations of multimodal, heavily skewed distributions are not good representations of the underlying data and we were interested in characteristic modes of the distributions, we used mixture models to identify dominant modes of the posterior distributions.

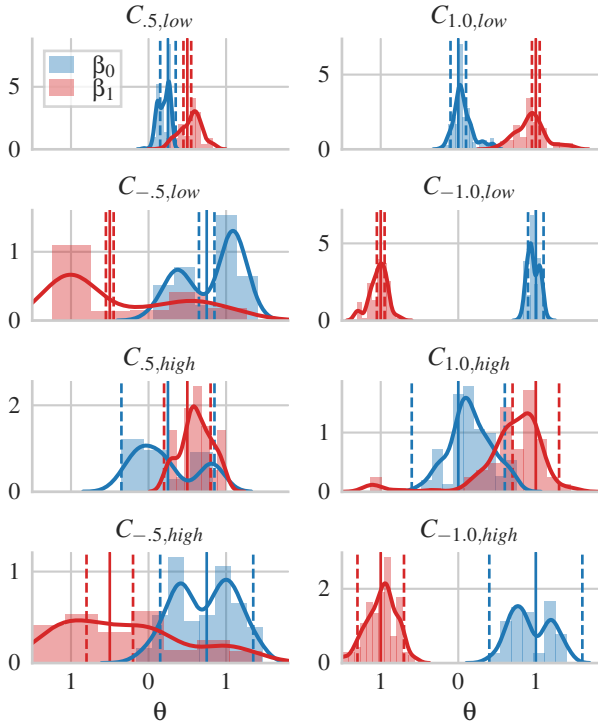

Figure 4: Posterior densities for one participant in each condition. Lines represent the true values and standard deviations (dashed lines) in the experimental conditions.

Table 3: Posterior means and variances per condition, for function intercepts (β_0) and slopes (β_1).

Condition	M_{β_0}	SD_{β_0}	M_{β_1}	SD_{β_1}
$C_{.5,\text{low}}$	0.34	0.32	0.32	0.62
$C_{1.0,\text{low}}$	0.52	0.40	0.00	0.78
$C_{-.5,\text{low}}$	0.49	0.37	−0.02	0.73
$C_{-1.0,\text{low}}$	0.65	0.40	−0.40	0.77
$C_{.5,\text{high}}$	0.35	0.39	0.27	0.71
$C_{1.0,\text{high}}$	0.47	0.40	0.07	0.81
$C_{-.5,\text{high}}$	0.54	0.40	−0.07	0.77
$C_{-1.0,\text{high}}$	0.52	0.44	−0.20	0.83

Estimating Posterior Density Clusters We estimated Gaussian mixture models that best described the distributions for each experimental condition. We incrementally increased the number of components and selected the model with the lowest BIC[4]. The clustering produced a moderate number of clusters, reflecting the multimodal nature of the data. In general, each condition was estimated to correspond to a mixture of 1–8 clusters ($M = 4.5, SD = 2.56$), and the largest clusters closely matched the different training conditions. For KL-divergences between training distribution and the inferred clusters, see Table 5, for the number of clusters, weights, means and covariances for the largest clusters, see Table 4, for plots of the clusters, see Figure 5.

[4]Estimating the mixtures with a Bayesian Dirichlet process mixture model yielded very similar results.

Table 4: The total number of clusters (N_c) assigned was generally low and the weight of the largest clusters was relatively large (16–100%).

Condition	N_C	$w_{c=1}$	$\mu_{\beta_{0,c=1}}$	$SD_{\beta_{0,c=1}}$	$\mu_{\beta_{1,c=1}}$	$SD_{\beta_{1,c=1}}$
$C_{.5,\text{low}}$	8	0.2	0.15	0.02	0.69	0.14
$C_{1.0,\text{low}}$	8	0.17	0.07	0.01	0.84	0.1
$C_{-.5,\text{low}}$	1	1.0	0.49	0.14	-0.01	0.53
$C_{-1.0,\text{low}}$	4	0.42	0.93	0.03	-0.98	0.04
$C_{.5,\text{high}}$	2	0.81	0.24	0.10	0.54	0.21
$C_{1.0,\text{high}}$	3	0.46	0.24	0.1	0.75	0.13
$C_{-.5,\text{high}}$	5	0.31	0.93	0.09	-0.65	0.2
$C_{-1.0,\text{high}}$	5	0.39	0.9	0.08	-0.95	0.07

Table 5: KL-divergence between the training distribution and the three largest clusters. In general, one of the largest clusters corresponded well to the training distribution.

Condition	$KL_{c=1}$	$KL_{c=2}$	$KL_{c=3}$
$C_{.5,\text{low}}$	2.18	1.1	1.95
$C_{1.0,\text{low}}$	1.74	42.1	5.85
$C_{1.0,\text{low}}$	1.35	–	–
$C_{-1.0,\text{low}}$	0.31	1.76	24.16
$C_{.5,\text{high}}$	0.83	10.37	–
$C_{1.0,\text{high}}$	1.06	9.76	2.95
$C_{-.5,\text{high}}$	1.49	2.66	4.25
$C_{-1.0,\text{high}}$	1.02	59.27	11.09

Per-Participant Clusters To evaluate if the source of the multimodality in our data was due to averaging over diverse cohorts of participants, or if individual participants produced multimodal posteriors, we performed the same clustering procedure on a per-participant basis. Participant posterior distributions were characterized by 1–12 clusters ($M = 3.11, SD = 1.96, Q1 = 1, Q2 = 3, Q3 = 4$), suggesting that the posterior distributions were composed of multimodal individual distributions. Furthermore, some participants with optimal $\hat{R}$ (≤ 1.1) also exhibited multiple clusters, indicating that the multimodality was not simply due to poor convergence ($M = 1.89, SD = 1.36, N_{\hat{R}\leq 1.1} = 9$).

The number of clusters did not differ significantly between low- and high-variance conditions, $M_{low} = 2.98, SD_{low} = 1.94, M_{high} = 3.1, SD_{high} = 1.57, t(164.24) = -0.49, p = .312$. Neither did the variance of the largest cluster for slopes differ significantly, $M_{low} = 0.1, SD_{low} = 0.13, M_{high} = 0.1, SD_{high} = 0.11, t(172.43) = 0.11, p = .545$. However, for intercepts the variance of the largest clusters was significantly different, with smaller cluster variances for low-variance conditions, $M_{low} = 0.04, SD_{low} = 0.03, M_{high} = 0.05, SD_{high} = 0.04, t(189.85) = -2.09, p = .048$.

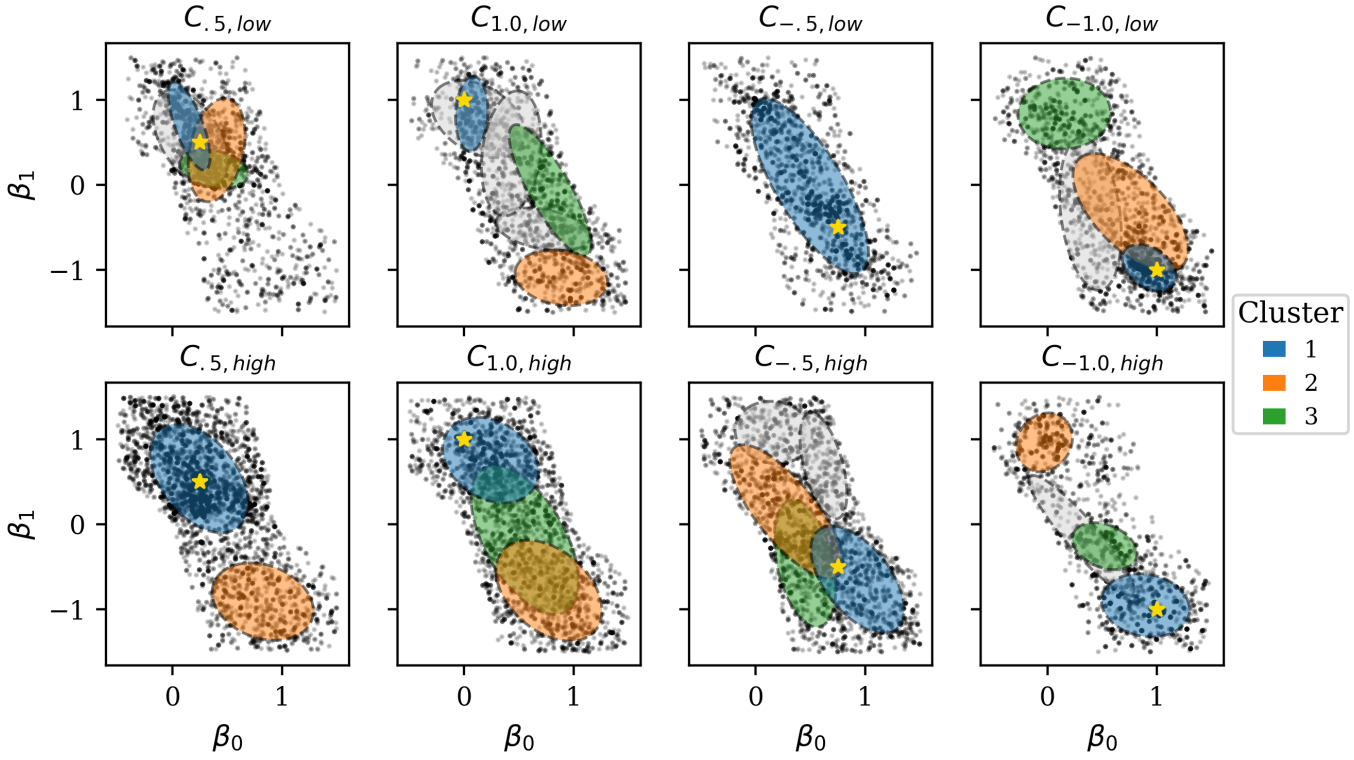

Figure 5: Clusters obtained by fitting a Gaussian mixture model (oval shapes). The top three clusters (colored shapes) accounted for a large proportion of the data and in general matched the distribution learned in the training phase well (mean parameters of the true distribution in yellow).

Discussion

We have found some evidence that participants represent the functions learned in training as distributions over parameters. Furthermore, the modes of these distributions were, in many cases, aligned with the true parameters. In addition, for intercepts, but not for slopes, these distributions were affected by differences in the variability of training. Finally, our results suggest that the learned distributional spaces over function parameters can exhibit multiple modes.

The multimodality in the posterior distributions allows for two interpretations. First, it is possible that participants truly evaluated distinct candidate representations, and thus multimodal posterior distributions characterized their hypothesis space. It is plausible that highly salient relationships, in addition to the implied parameters in the training, constitute the psychological space when learning sets of varying functions. However, the multimodality might also arise from our experimental method. One issue could be the number of iterations. Theoretically, MCMCP is well suited to discover complex, multimodal distributions, but practically many more samples could be necessary to achieve convergence to the posterior distribution. Since extremely large numbers of iterations might not be feasible from an experimental perspective, one practical test of our results could be starting the chains of later participants at the endpoints of previous participants (Martin et al., 2012).

Future research should clarify the source of multimodality, for instance by comparing our results with results obtained by multidimensional scaling (MDS). If such a comparison corroborates our results, these insights into the structure of psychological spaces could, in turn, provide invaluable guidance for future generalization research. In addition, MDS would also allow us to address two shortcomings of the current study: its exclusive focus on linear functions, and the potential influence of perceptual similarity of functions on participants' forced choices. First, similarity judgments obtained via MDS could be used to determine if participants are well described by linear models, or if non-linear representations underlie their judgments. These results would allow us to determine if the multimodal representations observed in our experiment were the result of a lack of satisfactory choices or a genuine characteristic of learning. Second, MDS would allow us to chart sets of perceptually similar samples. It is plausible that intercepts and slopes can affect notions of similarity of linear functions differently. For example, if functions sharing the same slope but very different intercepts are judged more similar than functions with similar slopes and intercepts, such non-linear interactions could explain the multimodality observed in our experiment.

While more research is required, our results also highlight the importance of a plurality of experimental approaches and methods in the study of human generalization. Most of previ-

ous research has focused on averaged errors or single extrapolations. Here, we suggest that to fully understand human generalization, characteristic errors, in combination with extrapolation patterns, and evaluation and exploration of the underlying hypothesis spaces are required.

Acknowledgements

We thank Yevgen Matusevych, Arabella Sinclair and three anonymous reviewers for their helpful comments and suggestions. ANS was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology*, *30*(1), 38.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Brown, M., & Lacroix, G. (2017). Underestimation in linear function learning: Anchoring to zero or xy similarity? *Canadian Journal of Experimental Psychology*, *71*(4), 274–282.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology*, *23*(4), 968.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology*, *32*(5), 1019.

León-Villagrá, P., Preda, I., & Lucas, C. G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Lucas, C. G., Sterling, D., & Kemp, C. (2012). Superspace extrapolation reveals inductive biases in function learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of markov chain monte carlo with people using facial affect categories. *Cognitive science*, *36*(1), 150–162.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

Ramlee, F., Sanborn, A. N., & Tang, N. K. (2017). What sways peoples judgment of sleep quality? A quantitative choice-making study with good and poor sleepers. *Sleep*, *40*(7).

Roberts, G., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive psychology*, *60*(2), 63–106.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. (2015). Assessing the perceived predictability of functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Stojic, H., Eldar, E., Hassan, B., Dayan, P., & Dolan, R. J. (2018). Are you sure about that? On the origins of confidence in concept learning. In *Preceedings of the Annual Conference on Cognitive Computational Neuroscience.*

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*.

Wilson, A. G., Dann, C., Lucas, C. G., & Xing, E. P. (2015). The human kernel. In *Advances in Neural Information Processing Systems* (pp. 2854–2862).