Simplicity Bias in the Estimation of Causal Functions

Daniel R. Little (drlittle@indiana.edu)
Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St, Bloominton IN 47405 USA

Abstract

We ask observers to make judgments of the best causal functions underlying noisy test data. This method allows us to examine how people combine existing biases about causal relations with new information (the noisy data). Participants are shown n data points representing a sample of noisy data from a supposed experiment. They generate points on what they believe to be the true causal function. The presented functions vary in noise, gaps, and functional form. The method is similar to function learning studies, but minimizes the roles of learning and memory. To what degree do the participants exhibit a bias for simple linear functions? We describe a hierarchical Bayesian polynomial regression model to quantify complexity. The results show the expected bias for simplicity, but with some interesting individual differences.

Keywords: causal models, function learning; prior knowledge; hierarchical Bayesian regression.

In science generally, in statistical inference, in studies of model selection, and in psychological theorizing, we design our inference approaches to trade off fit to observed data (models are good that fit well) and complexity (models or explanations that fit or explain everything are bad). In the present research we explore how observers form causal models for noisy data by asking observers to estimate functions as explanations for a set of noisy data points. We ask: How do mental causal models balance fit and complexity? Further, how does the balance of fit and complexity achieved by observers compare with the balance imposed by a rational system? We explore the former question by fitting Bayesian hierarchical models to the causal functions produced by the observers. The latter question is addressed by comparing the model for the observers to a Bayesian hierarchical model fit to the presented noisy data.

Recent research into beliefs about casual relationships has demonstrated that people prefer simple explanations over more complex explanations (see e.g., Lombrozo, 2006, 2007). In fact, simplicity biases have been forwarded as underlying many fundamental cognitive processes (Chater & Vitanyi, 2003). Here we are primarily concerned with how people form mental models summarizing relationships between continuous input and output quantities (i.e., data). Learning about continuous variables has typically been studied as a *function learning* phenomena. Within this paradigm, input values from a single function are displayed one at a time, the participant responds with an output value and receives corrective feedback about the true output value. Over time, participants form some representation about the underlying functional relationship between the two variables (i.e., exemplar-based associations, DeLosh, Busemeyer, & McDaniel, 1997; or a mixture of linear function 'experts', Kalish, Lewandowsky, & Kruschke, 2004). Typically, the focus in function learning is on the type and relative difficulty of learning different functions; most function learning studies have not directly examined prior beliefs. However, there have been some recent investigations of function biases in the iterated learning paradigm (see e.g., Griffiths, Kalish, & Lewandowsky, 2008; Kalish, Griffiths, & Lewandowsky, 2007). In an iterated learning task, participants are trained on stimuli drawn from the learning outcomes of the previous participant. Each learner is assumed to have a set of hypotheses about the causal relationship between the input and output variables. A rational learner assigns probabilities to each hypothesis in accordance with the posterior probability of that hypothesis given the data. According to Bayes' rule the posterior probability, $p(h|d)$, is:

$$p(h|d) = \frac{p(d|h)p(h)}{p(d)} \tag{1}$$

where $p(d|h)$ is the probability or likelihood of the data given a hypothesis, $p(h)$ is the prior probability of that hypothesis and $p(d)$ is the marginal probability of the data, $\sum_i p(d|h_i)p(h_i)$. In an iterated function learning task, responses tend to converge to a positive linear function indicating a bias towards positive linear functions in people's prior beliefs about functions(Kalish et al., 2007).

Non-Bayesian approaches to function learning also provide hints of a simplicity bias. Humans learn positive linear functions faster than negative linear functions or nonlinear functions (Carroll, 1963). One successful theoretical approach assumes that performance is driven by combining a pre-existing set of linear functions (i.e., POLE; Kalish et al., 2004). However, the representations formed in function learning studies are determined in part by cognitive limitations on, and the processes of, attention, learning, and memory. For example, the relation between an input value and output value presented at one point in time might be partially or wholly forgotten later, and possibly distorted through inference in the direction of simplicity. As another example, extreme points might be attended and remembered best, leading to a different set of distortions. In the current paper, we present an experimental method for studying function representations in which all the data is presented simultaneously, in a form easy to perceive and process. In this method, one should not have to parcel out the effects due to faulty memory for the data points on the function, or inadequate learning of the function over time. We use the method to address two related questions: do people show a bias towards simple (possibly linear) functional relationships when faced with

noisy data generated from a range of different function types? and b) how do people's prior beliefs about functional relationships interact with information given to them about the function? Let h represent a causal functional form. We assume that people are rational to the degree that they properly combine prior expectations, $p(h)$, with information from the problem, $p(d|h)$. In our modeling, we will estimate $p(h)$, and the distribution of likely values of $p(h)$, from the responses. We assume that the distribution of of $p(h)$ is a direct measurement of the prior biases.

A Method for Studying Function Biases

In a typical function learning task, participants are shown values from a *single* function, values are presented one at a time and participants only learn about one function in the course of the experiment. In order to separate the application of functional knowledge from learning and memory, the current experimental method showed participants a selection of data points all generated by the same function and presented simultaneously (with Gaussian noise added to obscure the true function; see Figure 1). The participants were asked to demonstrate their best estimate of the true underlying function by placing a series of points at specified places in the presented graph. They were told that these response points should lie on their best estimate of the underlying causal function for the data on the current trial. The data on each trial were drawn from different functions and across trials the functions varied in number of data points that were displayed and the amount of Gaussian noise added to the output.

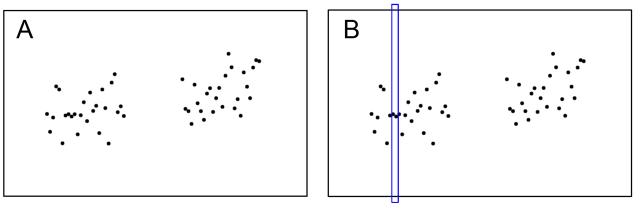

Figure 1: A: Example of "data" presented to participants during the function estimation task. B: Example of the "data" presented to participants along with a response column. The participant places a 'best' point with the response column. Many such columns are presented on each trial.

Bayesian polynomial regression

In a *regression* problem, the goal is to describe the functional relationship between two continuous variables. In Bayesian terms, we need to compute the posterior probability of each possible function given the data. The hierarchical approach allows us to specify probability distributions to capture our uncertainty about not only the best-fitting parameter values within a model class (e.g., the slope and intercept of a linear function) but also at higher levels of abstraction, such as the entire set of model classes (all polynomials). Using Markov Chain Monte Carlo (MCMC) techniques we can sample from the distributions at higher levels. For example, if we specify a probability distribution over polynomial degree, we can estimate the shape of this distribution using MCMC; hence, any bias towards using less complex functions is reflected by increased mass over lower degree polynomials.

The first step in hierarchical regression is to specify the probability distributions over the parameters within a model class to a given set of data (e.g., a single trial j from the current experiment). If we assume that people explicitly represent functions, then the likelihood of the observed data given a specific model class k is given by:

$$p\left(y|\beta,\sigma^2,X_k\right) \sim N\left(X_k\beta,\sigma^2 I\right), \tag{2}$$

where β is the vector of regression coefficients, X_k is the matrix of predictor variables for model k, and $\sim$ means "is distributed as" (in this case, the distribution is a multivariate normal distribution with a specified mean and covariance matrix). Here we consider the models to be polynomials of the form $y = \sum_{i=0}^{k} \beta_i x^i + \varepsilon$ where $\varepsilon \sim N\left(0,\sigma^2\right)$.[1]

Within each particular model class k, if we assume noninformative priors over the coefficients, β, and the variance, σ^2, then the posterior probability of the coefficients is given by:

$$\beta \sim N\left(\hat{\beta}\mathbf{X},\mathbf{V}_\beta\boldsymbol{\sigma^2}\right), \tag{3}$$

where $\hat{\beta} = \left(\mathbf{X^T X}\right)^{-1}\mathbf{X^T y}$ and $V_\beta = \left(\mathbf{X^T X}\right)^{-1}$. The posterior probability of the variance is:

$$\sigma^2|y \sim Inv-\chi^2\left(n-k,s^2\right), \tag{4}$$

where n is the number of data points, k is the degree of polynomial under consideration, and $s^2 = \frac{1}{n-k}\left(y-X\hat{\beta}\right)^T\left(y-X\hat{\beta}\right)$. Using noninformative priors can result in improper posterior distributions unless a) $k < n$ and b) the matrix of predictors, X, is full rank (Gelman, Carlin, Stern, & Rubin, 2003). The above distributions describe how to determine the posterior probabilities for the parameters of a given model class given one set of data; the focus of our analysis, however, is on how people choose amongst model classes over the entire set of experimental data. Hence, the main quantity of interest is the marginal distribution of the data for each model class k (the probability of the data over all parameter settings within each model, $p(y|k) = \sum_i p(y|\beta_{k_i},\sigma^2_{k_i})p(\beta_{k_i},\sigma^2_{k_i})$ for each problem j weighted by the prior probability of selecting k for each particular problem).

The marginal probability of the data is used to compute the posterior probability that the model class is k by using Equation 1; hence, we also need to define a prior over polynomial degrees representing the probability of selecting model k for any problem j. We assume a prior distribution over the model class k given by:

[1]The data, matrix of predictors, the coefficients, the variance, and the number of data points, n, change across trials, j, but for notational simplicity we suppress indexing by trial.

$$k \sim Categorical\,(p_1, ..., p_K)\,. \tag{5}$$

Polynomial degree is thus treated as a categorical variable with the probability of selecting a particular degree given by the categorical distribution parameters, $p_1, ..., p_K$. Intuitively, these parameters index prior belief about model classes of different complexity. If we were solely concerned with selecting the appropriate model for the data, we could set the multinomial parameters to reflect our assumption of a simplicity bias over the model class (i.e., $p_1 > p_2 > ... > p_K$). However, we want to measure prior bias without committing ourselves to any *a priori* assumptions about simplicity; hence, a non-informative prior distribution over the multinomial parameter must also be specified. We use the conjugate prior to the categorical as follows:

$$p_1, ...p_K \sim Dirichlet\,(\alpha_1, ..., \alpha_K) \tag{6}$$

where K is the number of different polynomial degrees being tested and $\alpha_1, ..., \alpha_K = 1$, which is equivalent to a multivariate uniform distribution. MCMC techniques including Gibbs Sampling (see e.g., Gelman et al., 2003) allow the prior distribution of the model degree k to be approximated.[2]

This analysis when conducted on each participant's responses gives a direct measure of the prior expectation for polynomials of different complexity in the distribution of $p_1, ...p_K$. This analysis also provides the posterior distribution of the degrees, (hereafter, $k_{response}$), for each problem. When conducted on the data that are shown to the participants the distribution of the degrees, (hereafter, k_{data}) gives a measure of optimal model selection for each problem (i.e., the degree k of the model that best captures the data). By comparing the data and response distributions of k for each problem, we can investigate how information from the problem is combined with prior biases.

Method

Participants and Design

Five Indiana University psychology graduate students participated in the experiment and received $16 dollars reimbursement. Each participant responded to 108 functions which varied in a) the number of data points displayed on screen (6, 16, or 52 points), b) the variance of the Gaussian noise $\left(\mu = 0, \sigma^2 \in (.25, 1)\right)$ added to each y-value, and c) the generating function, which was either a polynomial of 1, 2, 3, 4 or 5 degrees (i.e., linear, quadratic, cubic, etc), or a sin, tan or exponential function. Function coefficients were uniformly generated with replacement from the integers 1 to 10. Different random coefficients and Gaussian noise were generated for each participant; hence, each participant saw a different instantiation of the base functions. [3]

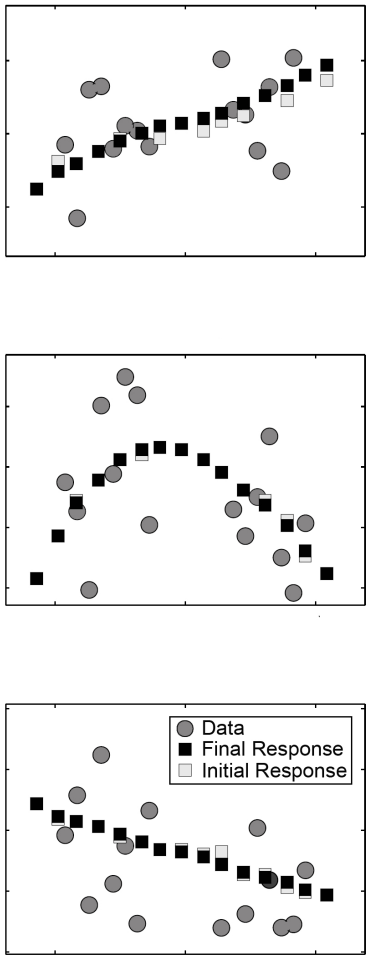

Figure 2: Example of data and responses from three different trials for one participant.

Procedure

On each trial, a data set (e.g., see Figure 1) was presented on the screen, and participant estimated the function that best summarized the causal relationship between the input and output values. Their response function was queried by prompts at several locations on the screen. There were 15 successive response requests. Each highlighted an x-value for the entire height of the screen (see Figure 1). Participants responded with their function value for a given x-value by clicking on their best guess location in the response column

[2] Here we use the *pseudo-prior* sampling method from Carlin and Chib (1995) to ensure that the sampling method switches between models. The parameter distributions for β are updated from Equation 3 for each model m when the selected model $k \neq m$. When $k = m$, β is sampled from the distribution in Equation 3 but with the variance in that distribution multiplied by C, where $C << 1$; hence, the distribution of the coefficients is more diffuse for model j when $k = j$.

[3] For each participant, each function was presented three times. Participants tended to respond with functions of the same complexity across repetitions. However, there was variability in the actual function generated, and we therefore do not aggregate across repetitions. The proportion of response functions that differed in their polynomial degree across with problem repetitions was .17, .06, .13, .06, and .06 for each participant, respectively.

with the mouse. During the initial response stage, the selected function value was removed until all of the response columns had been presented. After responding to the 15 response columns, all of the participant's initial responses were presented simultaneously on the screen, and participants had the opportunity to adjust their functions until they were satisfied with the result.

Results

An example of the data shown to the participants, the initial function response and the final function response are shown in Figure 2. We used the Bayesian regression model outlined above to estimate the complexity (i.e., the degree of the polynomial with the highest mean posterior probability) of the data and the complexity of each participant's final response function. A scatterplot of the estimated polynomial degrees for the data and the responses are shown in Figure 3 (top row).

The hierarchical Bayesian regression allows examination of the prior density over the polynomial degree for each participant (that is, $p_1, ..., p_k$ for all degrees k). The prior densities for each participant are shown in Figure 3 (bottom row). These densities clearly show that all of the participants demonstrate a bias towards functions of lower polynomial degree. Three of the participants show a higher mass over linear functions, while two of the participants show a higher mass over horizontal functions. Consideration of the Bayesian solution (see the top row of in 3) indicates that the a bias towards simple functions is appropriate for this task. The Bayesian solution also prefers simple functions for many of the problems. We consider the relationship between the optimal function and the generated function in more detail below.

Combining data with prior expectations

To examine the factors that influenced responding, we used the hierarchical Bayesian regression model to predict the mean polynomial degree of responses, $k_{response}$, using the following predictors: a) The mean polynomial degree of the *data*, k_{data} (i.e., the complexity of the data that participants were shown). b) An estimate of the noise (σ_P^2) generated using Equation 4 with $\hat{\beta}$ and X estimated from the *response* function. c) The number of data points, N. In addition, the pairwise interactions between k_{data}, σ_P^2, and N, the three-way interaction between all variables plus an intercept. For this analysis we are only concerned with the posterior probabilities of the coefficients; hence, we integrate out σ^2 and compute the posterior probability of the coefficients as a multivariate t-distribution, $\beta|k_{response} \sim t_{n-k}\left(\hat{\beta}\mathbf{X}, \mathbf{V}_\beta\boldsymbol{\sigma^2}\right)$. The posterior distributions of the coefficients for each participant are shown in Figure 4. Coefficients whose 95% credible intervals do not overlap zero are shown in bold.

Clearly, all of the participants responses are predicted by the intercept (which provides a measure of the base polynomial degree) and the complexity in the data, k_{data}. For two of the participants (numbers 3 and 4; see rows C and D in Figure 4), only these two variables predict $k_{response}$. For these two participants, the average intercept value is less than 1, which corresponds to the prior distribution of $k_{response}$ shown in Figure 3 (bottom row, columns 3 and 4). The means of the k_{data} distributions are also less than one indicating that for these two participants an increase in the complexity of the data is accompanied by an lesser increase in the complexity of the response. For the other three participants, the influence of the intercept and the complexity of the data are augmented by other factors. For participant #2, the coefficient for the estimated noise, σ_P^2, is also greater than zero indicating a tendency to respond with higher $k_{response}$ with noisier data (see Figure 4, row B). For this participant, the estimated intercept value also accords well with the prior distribution shown in Figure 3 (bottom row, column 2). Responses from participants 1 and 5, in addition to the intercept and k_{data} variables, are also influenced by the interaction between the complexity of the data and the number of data points, $k_{data} \times N$, revealing a tendency to increase the complexity of their responses for complex data with higher N (see Figure 4, rows A and E). Participant #1 also showed a tendency to increase response complexity for noisy complex functions (see Figure 4, row A).

Discussion

The posterior distributions over $k_{responses}$ clearly show a bias toward simple functions. This bias is warranted by posterior distribution over k_{data}. This finding accords well with results from related domains using a single function type such as iterated learning or function learning.

Not surprisingly, all participants showed a tendency to increase the complexity of their responses whenever the complexity of the data increased. This combination of data and prior follows directly from Bayes' rule (see Equation 1); in the order for the posterior distribution over $k_{response}$ to shift towards higher degree polynomials, the probability of the data given higher degree polynomials must be high enough to outweigh the prior bias towards lower degree polynomials. It is reasonable to assume that this only occurs when the data are complex, and the results seem to support this notion.

The Bayesian analysis also offered a straightforward way to handle individual differences in not only the prior distributions over $k_{response}$ but also in how information is extracted about the problem with some participants also utilizing the number of data points and the noise in the data in producing their responses. One limitation of the analysis presented here is that the polynomial regression returns a complete distribution over k for each problem; however, the follow-up regression analysis only utilizes the mean k. Future analysis would combine the regression analyses by making the polynomial degree of the responses contingent on the outcome of the regression model used in the follow-up analysis.

In typical function learning studies, the processes and limitations associated with attention, learning, and memory can obscure the underlying way that observers trade fit and complexity in their formation of mental causal models. By con-

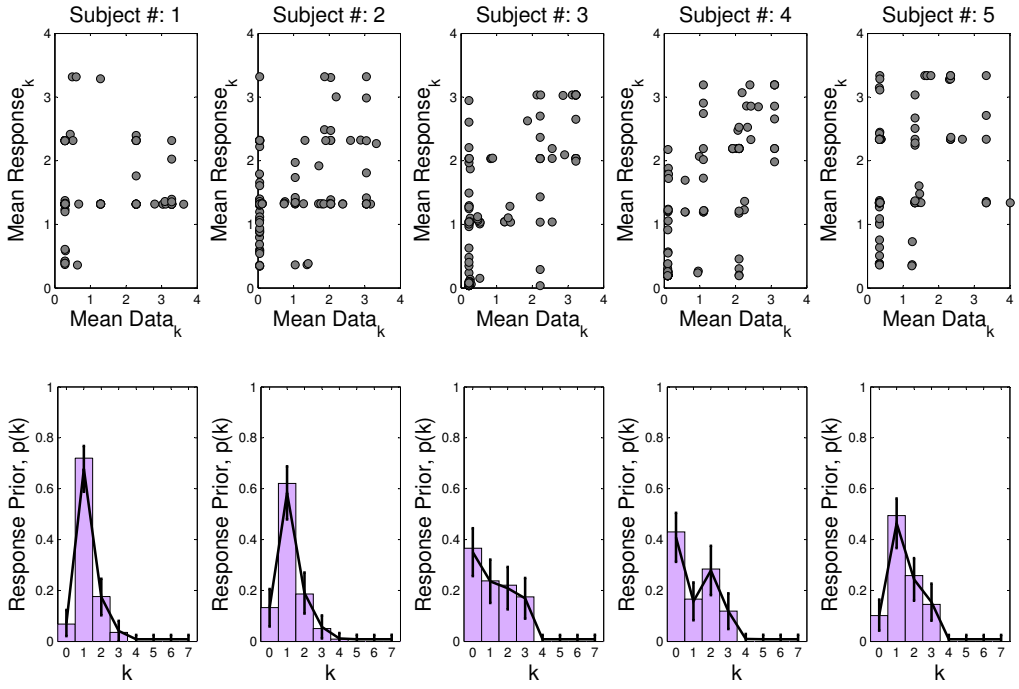

Figure 3: Top Row: Average estimated degree of best fitting polynomial for the data plotted against the average best fitting polynomial for the responses. The data points have been jittered to reduce overlap. Bottom Row: Histogram the sampling distribution of the prior density of polynomial degrees for each participants responses ($p(k)$). The estimated mean and 95% credible intervals are also shown.

trast, the presented methodology offers several advantages: 1) the influence of memory and learning on responses is reduced and 2) data from a large number of functions that differ in type and complexity can be collected efficiently. 3) Additionally, important factors, such as noise and the amount of data, can be easily manipulated without any detriment in performance. In function learning, people are not very good at learning about noisy functions (see e.g., Carroll, 1963). These three points have particular value when studying the prior biases underlying function representation. However, there are several other findings from function learning which are not addressed here. For instance, extrapolation tends to be linear and is an important diagnostic result (see DeLosh et al., 1997); only a small extrapolation region was included in the current experiment (see Figure 1). Future work will increase the number of responses collected from the extrapolation region to allow examination of extrapolation biases when learning and memory are removed from information processing.

Finally, one clear divergence in the present work from previous work is the use of a polynomial regression model to capture function-based performance. While early work in function learning assumed that people explicitly represented the to-be-learned functions (Carroll, 1963), more recent work has emphasized similarity-based processing. The basic idea is that each input that is presented during learning is stored in memory along with an associated output, new inputs are compared to the stored values and responses are generalized based on the similarity (i.e., psychological distance) between the new value and all of the stored values (DeLosh et al., 1997). Griffiths, Lucas, Williams, and Kalish (in press) have recently demonstrated that the function-based view and the similarity-based view are related: Bayesian linear regression requires predicting y by specifying a prior over the class of polynomials; if we instead specify a covariance matrix on the different input values (e.g., by taking, for example, the exponential of the squared distance between all pairwise combinations of x-values) then Gaussian process regression achieves the same result. Gaussian process regression is essentially a similarity-based view of function learning that is isomorphic to Bayesian linear regression. The experimental method employed in this paper offers a promising way to explore both views of function-based performance.

Acknowledgments

The authors would like to thank Woojae Kim helpful comments on the computational analysis. Preparation of this article was facilitated by an NIH-NIMH training grant #:T32 MH019879-14 to the first author.

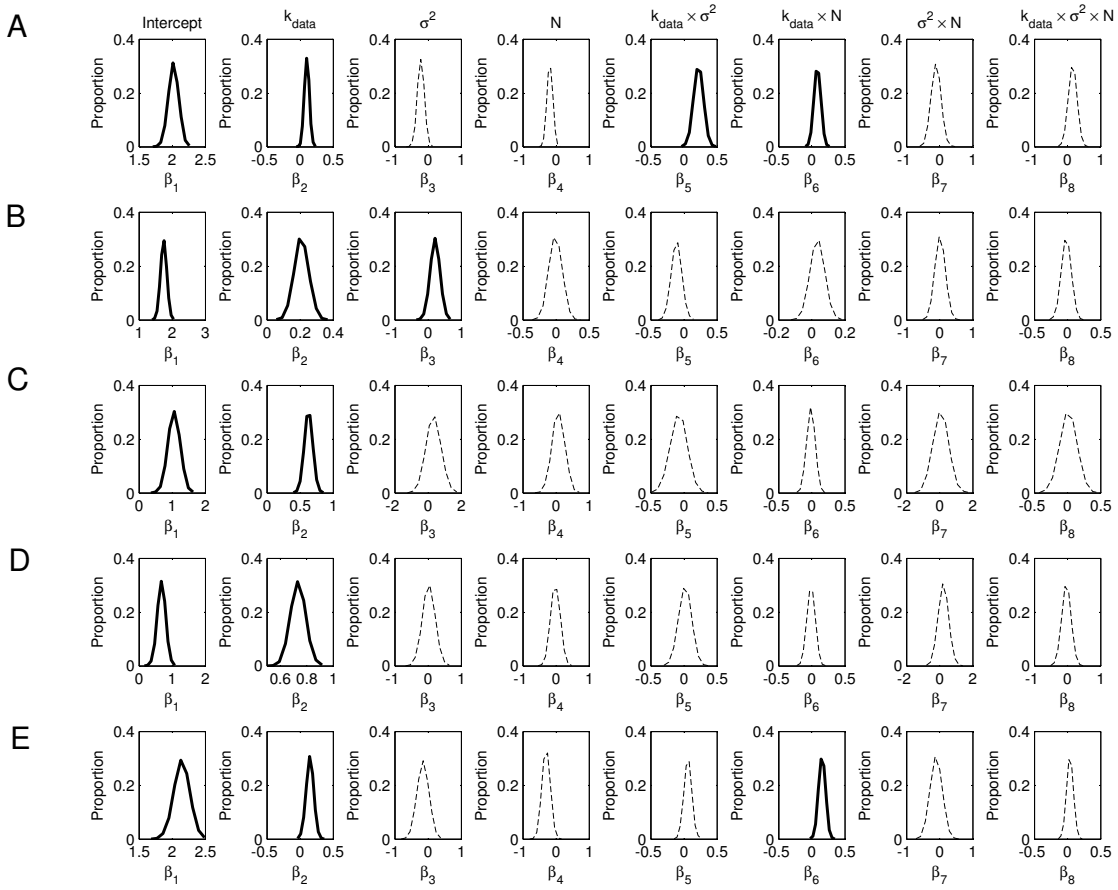

Figure 4: Distribution of regression coefficients for A: participant 1, B: participant 2, C: participant 3, D: participant 4, E: participant 5. Coefficients whose 95% credible intervals do not overlap zero are shown in bold.

References

Carlin, J. B., & Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, *57*, 473-484.

Carroll, J. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *Educational Testing Service Research Bulletin (RB-62-26)*.

Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19-22.

DeLosh, E., Busemeyer, J., & McDaniel, M. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *23*, 968-986.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall CRC.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transcripts of the Royal Society B*, *363*, 3503-3514.

Griffiths, T. L., Lucas, C., Williams, J. J., & Kalish, M. L. (in press). Modeling human function learning with gaussian processes. *Advances in Neural Information Processing Systems*, *21*.

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

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072-1099.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Science*, *10*, 464-472.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232-254.