Assessing the Perceived Predictability of Functions

Eric Schulz[1](e.schulz@cs.ucl.ac.uk), **Joshua B. Tenenbaum**[2](jbt@mit.edu), **David N. Reshef**[3](dnreshef@mit.edu),
Maarten Speekenbrink[1](m.speekenbrink@ucl.ac.uk), & **Samuel J. Gershman**[2](sjgershm@mit.edu)
[1]Department of Experimental Psychology, University College London, London, WC1H 0AP
[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA
[3]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, USA

Abstract

How do we perceive the predictability of functions? We derive a rational measure of a function's predictability based on Gaussian process learning curves. Using this measure, we show that the smoothness of a function can be more important to predictability judgments than the variance of additive noise or the number of samples. These patterns can be captured well by the learning curve for Gaussian process regression, which in turn crucially depends on the eigenvalue spectrum of the covariance function. Using approximate bounds on the learning curve, we model participants' predictability judgments about sampled functions and find that smoothness is indeed a better predictor for perceived predictability than both the variance and the sample size. This means that it can sometimes be preferable to learn about noisy but smooth functions instead of deterministic complex ones.

Keywords: Function learning, predictability, smoothness, Gaussian processes

Introduction

Learning about functions from noisy observations is a ubiquitous task. How much food should I cook to satisfy every guest at a party? How much do I have to turn the faucet handle in order to get the right temperature? An important problem facing a learner in the real world is choosing what functions to learn. While the number of functional relations between variables is virtually limitless, only predictable functions are worth learning. For example, one might attempt to learn a function relating gas prices to batting averages, but this function is unlikely to be predictable in the sense that the learned function can generalize accurately to new inputs. In this paper, we provide a formal framework for predictability, and study its implications for human function learning.

To appreciate why judging predictability is difficult, consider two kinds of functions: one that is complex (non-smooth) but nearly deterministic, and another that is simple (smooth) but noisy. Which function is more predictable? The answer is not obvious, but we show experimentally that human judgments exhibit a systematic preference in accordance with our theoretical analysis. Specifically, we find that—both theoretically and empirically—smoothness is a stronger determinant of predictability than noisiness.

To arrive at this result, we adopt a rational theory of function learning based on Gaussian process (GP) regression (Rasmussen & Williams, 2006). This theory unifies a number of earlier accounts (Koh & Meyer, 1991; DeLosh et al., 1997; McDaniel & Busemeyer, 2005), and provides a good fit to human function learning data (Griffiths et al., 2009). GP regression also lends itself to mathematical analysis, which we utilize in our modeling of predictability.

Background

Most previous research on human function learning has focused on interpolation and extrapolation (see McDaniel & Busemeyer, 2005, for a review). In an interpolation task, participants are presented with input-output pairs and then asked to make predictions about the outputs for test inputs that are between the training inputs. An extrapolation task is similar, but uses test inputs that are outside the convex hull of training inputs. Studies using these tasks have revealed what kinds of functions are easier to learn, and what kinds of inductive biases guide predictive judgments. For example, linear functions are usually easier to learn than non-linear functions, and non-monotonic functions are easier to learn than monotonic functions (Brehmer, 1974). People tend to exhibit a bias towards functions with positive linear slopes and an intercept of zero (Kwantes & Neal, 2006; Kalish et al., 2007). Other studies have shown that people are able to partition the input space into multiple distinct functions (Kalish et al., 2004).

A number of models have been proposed to account for these phenomena. Early theories posited the use of explicit rule-based functions (Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991), but these theories have trouble accounting for order-of-difficulty effects in interpolation tasks (McDaniel & Busemeyer, 2005), fail to predict extrapolation performance (DeLosh et al., 1997), and are unable to learn a partitioning of the input space (Kalish et al., 2004). Later theories used connectionist networks to capture many of these phenomena (DeLosh et al., 1997; Kalish et al., 2004; McDaniel & Busemeyer, 2005). In some cases (e.g., Kalish et al., 2004; McDaniel & Busemeyer, 2005) these networks incorporated rule-based functions into a hybrid architecture. One limitation of these theories is that they lack an obvious way to compute predictability. In the next section, we describe the GP theory of function learning (Griffiths et al., 2009), which offers a probabilistic perspective on predictability.

Gaussian process regression

Instead of assuming a parametric function class (as in early theories of function learning; Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991), GP regression places a prior distribution (namely, a GP) directly over the space of functions and carries out Bayesian inference in function space (Rasmussen & Williams, 2006). Let $f(x)$ be a function mapping an input x to an output y. A GP defines a distribution $p(f)$ over such functions. A GP is parametrized by a mean function $m(x)$ and a covariance function (or "kernel") $k(x, x')$:

$$m(x) = \mathbb{E}\left[f(x)\right] \tag{1}$$

$$k(x, x') = \mathbb{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] \tag{2}$$

Suppose we have observed n input-output pairs, $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = [x_1, \ldots, x_n]^\top$ and $\mathbf{y} = [y_1, \ldots, y_n]^\top$. We assume an additive noise model:

$$y = f(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

where σ^2 is the noise variance. Given a GP prior on the functions, $f \sim \mathrm{GP}(m, k)$, the posterior distribution over $f(x')$ given input x' is Gaussian with mean and variance given by:

$$\mathbb{E}[f(x')|\mathbf{x}, \mathbf{y}] = \mathbf{k}^\top (\mathbf{K} + \sigma^2 I)^{-1} \mathbf{y} \tag{4}$$

$$\mathbb{V}[f(x')|\mathbf{x}, \mathbf{y}] = k(x', x') - \mathbf{k}^\top (\mathbf{K} + \sigma^2 I)^{-1} \mathbf{k} \tag{5}$$

where $\mathbf{k} = [k(x_1, x'), \ldots, k(x_n, x')]^\top$ and $\mathbf{K}$ is the positive definite kernel matrix of covariances evaluated at the training inputs: $K_{ij} = k(x_i, x_j)$. The GP theory of function learning assumes that participants report $\mathbb{E}[f(x')]$ when asked to interpolate or extrapolate a function (Griffiths et al., 2009).

The covariance function encodes assumptions about what sorts of functions are probable *a priori* (i.e., it provides a form of inductive bias). Typically, this inductive bias corresponds to assumptions about the smoothness of functions over the input space, but assumptions about periodicity, linearity, and non-stationarity can also be encoded in the covariance function. These assumptions have important implications for learning and predictability, as we explore in the next section.

Learning curves

Theoretical learning curves relate the expected generalization error of a model to the amount of training data (Opper & Vivarelli, 1999; Williams & Vivarelli, 2000; Sollich & Halees, 2002). They can be seen as a mathematical expression of a function's predictability, given assumptions about the prior over functions, the noise process, and the distribution of inputs. Intuitively, a function exhibits a higher predictability if it is easier to predict new input points that are randomly chosen from the input space. If points are easier to predict, then the generalization error is lower as predictions will be closer to the underlying truth of the the actual function. Therefore, these two things, predictability and the generalization error, are two sides of the same coin (Goerg, 2013). The learning curves for GP regression can be used to derive *a priori* predictions about how different properties of functions such as smoothness, variance, and sample size influence perceived predictability. In particular, factors that increase the generalization error should lead to lower predictability judgments.

Given a dataset $\mathbf{x}$ and an error function $\mathcal{L}(\cdot, \cdot)$ which measures the difference between the predicted and true function values, the data-dependent generalization error is defined as the expected error on a test input x', marginalizing over the latent function:

$$E(\mathbf{x}) = \int_f p(f) \int_{x'} \mathcal{L}(f(x'), \bar{f}(x'))\, dx'\, df, \tag{6}$$

where $\bar{f}(x) = \mathbb{E}[f(x')|\mathbf{x}, \mathbf{y}]$. It is called the data-dependent error as it still depends on the position of the observed input points. The data-independent generalization error is defined as the expectation of $E(\mathbf{x})$ with respect to a density $p(\mathbf{x})$ on inputs with sample size n:

$$\mathcal{E}(n) = \int_{\mathbf{x}} p(\mathbf{x}) E(\mathbf{x})\, d\mathbf{x}. \tag{7}$$

It is called the data-independent error as it does not depend on the observations per se, but rather provides an a priori expectation of the error after n sample point have been observed. A learning curve is constructed by calculating the data-independent generalization error as a function of the sample size. While the learning curve is not analytically tractable (except for a few special cases), it is possible to derive a lower bound using the eigenfunction expansion of the covariance function: $k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x)$, where $\{\lambda_i\}$ is the spectrum of eigenvalues (decreasing as a function of i) and $\{\phi_i(x)\}$ are the eigenfunctions.

As shown by Opper & Vivarelli (1999), the generalization error for the squared-loss error function, $\mathcal{L}(y, \hat{y}) = |y - \hat{y}|^2$, can be lower-bounded by:

$$\mathcal{E}(n) \geq \sigma^2 \sum_{i=1}^{N} \frac{\lambda_i}{\sigma^2 + n\lambda_i}, \tag{8}$$

where N is the number of non-zero eigenvalues. Loosely speaking, the eigenvalue spectrum summarizes how the correlation between the function values of two points changes as a function of their input distance. Smoother functions have eigenvalues that decay more slowly across the spectrum. Smooth functions have long-distance correlations, which makes it easier to learn and therefore leads to smaller generalization errors.

The theoretical predictions of learning curves can be seen even more clearly when we look at covariance func-

tions with power-law spectral decay[1] (Sollich & Halees, 2002): $\lambda_i \propto i^{-r}$. Asymptotically ($\mathcal{E}(n) \ll \sigma^2$), the learning curve scales as

$$\mathcal{E}(n) \propto \left(\frac{\sigma^2}{n}\right)^{-\frac{r-1}{r}}. \quad (9)$$

This analysis tells us that the most important factor influencing a GP's learning curve is the smoothness of the covariance function (parametrized in terms of the spectral decay rate r). The generalization error depends polynomially on the variance but exponentially on smoothness. The implication is that noisy, smooth functions are more predictable than deterministic, complex functions. This is intuitive, because smooth functions allow data to be more strongly aggregated across different input points, whereas anything one can learn about a complex function is very local.

Parametrizing smoothness

To create functions with different levels of smoothness, we can employ a flexible class of stationary covariance functions constructed from the modified Bessel function (Rasmussen & Williams, 2006). Letting $\tau = |x - x'|$, the *Matérn* class of covariance function is given by:

$$k_s(\tau) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\tau}{\gamma}\right)^{\nu}\mathcal{K}_\nu\left(\frac{\sqrt{2\nu}\tau}{\gamma}\right), \quad (10)$$

where $\gamma > 0$ is a length-scale parameter, $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of order $\nu = s - 1/2$ for integral s, and $\Gamma(\cdot)$ is the gamma function. We will refer to s as the order of the Gaussian process.

When $s = 1$, Eq. 10 corresponds to the Ornstein-Uhlenbeck covariance function, which generates a process that is not mean square differentiable (see Rasmussen & Williams, 2006). This means that sampled functions will produce very rough outputs. When $s > 1$, the process is $s - 1$ times mean square differentiable, becoming smoother with increasing s. In the limit $s \to \infty$, it is equivalent to a squared exponential covariance function. We used Matérn functions of order up to 3 here as it is empirically hard to distinguish between functions of higher order. Figure 1 shows several sampled functions from Matérn covariance functions of different orders. It can clearly be seen how a higher s produces smoother functions.[2]

In the simple one-dimensional cases used here we can also easily simulate learning curves. This will provide us with a sanity check of the approximated learning curves derived above. Figure 2 shows the learning curves for

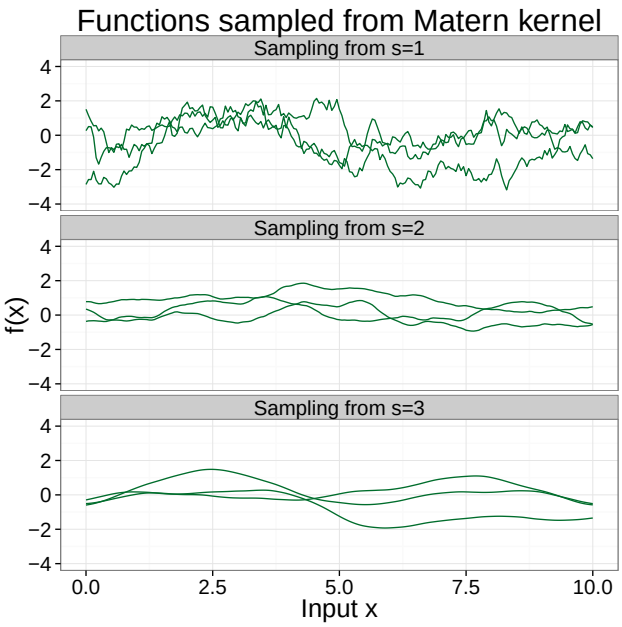

Figure 1: **Samples from GPs with Matérn covariance functions of different orders**.

two different covariance functions with two different degrees of added noise. The learning curves were derived by averaging 10,000 learning trials over 100 sequentially provided and evenly distributed input points.

Again, we can see that smoothness matters more than noise variance. However, there is a trade-off at the end of the scale as the added noise naturally defines the asymptote of the learning curves (if there is noise, one will always be a bit wrong). With these theoretical and simulation results in hand, we now turn to an experimental exploration of our formal account.

Experiment

We asked participants to judge the predictability of functions (displayed as a scatter plot), while manipulating the smoothness, noisiness and sample size. This allowed us to quantitatively measure the influence of these different factors on perceived predictability. Based on the analysis learning curves described in the previous section, we postulated the following 3 hypotheses:

1. Sample size and smoothness will correlate positively with perceived predictability, whereas noise variance will correlate negatively.

2. The effect of smoothness will be bigger than the effects of noise and sample size.

3. The approximate learning curve given a sample will be the most important factor influencing participants' predictability judgments overall.

[1] The one-dimensional Ornstein-Uhlenbeck covariance function described in the next section has this property, with $r = 2$.

[2] Further examples of functions are available at `http://bit.ly/1CtXfMA`.

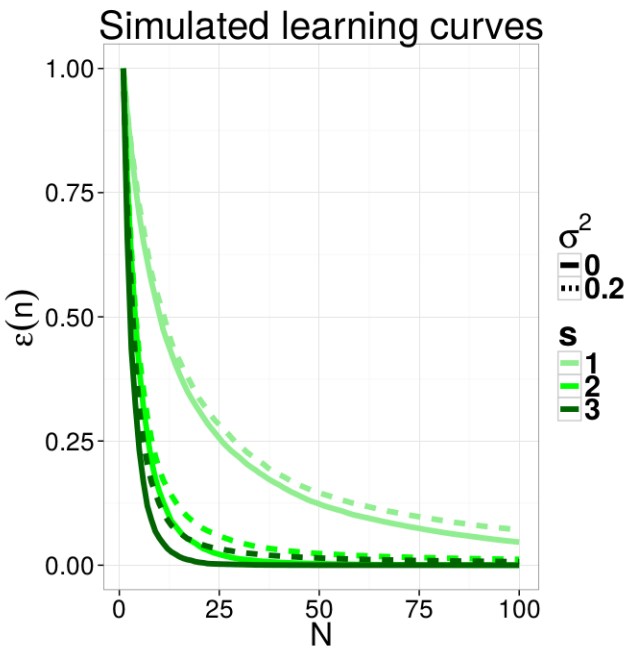

Figure 2: **Simulated learning curves**. Generalization error as a function of sample size for different levels of noisiness (σ^2) and smoothness (s).

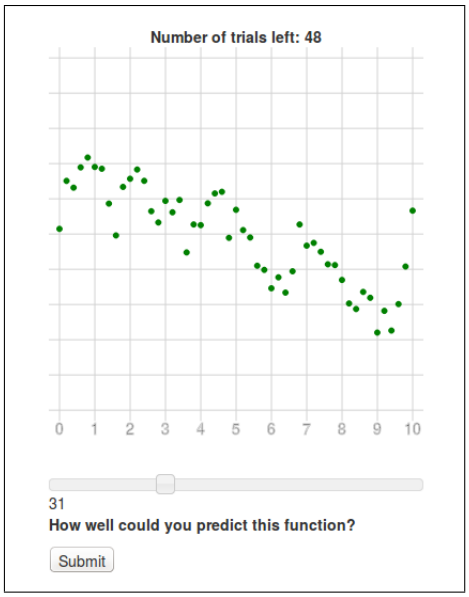

Figure 3: **Screenshot of experiment**.

Participants

47 participants were recruited via prolificacademic.co.uk and received £1 for their participation. 27 participants were female and the overall age had a mean of 24 with a standard deviation of 5.

Task

Participants were told that they had to assess how well they could potentially predict different functions on a scale from 0 (not at all) to 100 (certainly). It was explained to them that prediction means to assess a new input point uniformly sampled from the input range. They were sequentially shown 50 different samples of functions and had to indicate how well they thought they could predict a newly sampled point of that function. The functions were created online using Javascript.[3] A screenshot of the experiment is shown in Figure 3.

Design

Participants saw 50 trials where points were sampled equidistantly from different GPs with the Matérn covariance function. The parameters for the smoothness $s = [1, 2, 3]$, the variance $\sigma^2 = [0, 0.05, 0.1, 0.15, 0.2]$, and sample size $n = [10, 20, 30, 40, 50]$ were randomly selected on each trial. As GP samples were created on the spot, no participant saw the same function; only the different characteristics governing the generating process were manipulated. The length-scale of the covariance function was fixed at $\gamma = 1$.

[3]Code available at `github.com/ericschulz/gpsmooth`.

Results

Figure 4 shows the relationship between different GP parameters and perceived predictability. Increasing smoothness resulted in higher perceived predictability, and increasing noise variance reduced perceived predictability. The overall effect of an increasing sample size on the perceived predictability was negligible. This finding has at least two potential explanations: (1) It might be the case that participants overestimate the predictability with small sample sizes as they tend to infer smoother functions than the ones they actually see; (2) the equidistantly spaced inputs we presented to participants might permit easier prediction, since they cover more space overall.

	Estimate	**SD**	**t-value**	**Pr(>|t|)**
Intercept	41.70	1.82	22.93	0.00
n	0.79	1.02	0.77	0.44
s	8.00	0.72	11.11	0.00
σ^2	-5.21	0.66	-7.90	0.00
$n \times s$	1.48	0.38	3.90	0.00
$n \times \sigma^2$	-1.42	0.38	-3.73	0.00
$s \times \sigma^2$	-2.12	0.38	-5.54	0.00

Table 1: **Parameter estimates from mixed-effects regression analysis.**

We quantitatively assessed the influence of the different factors by performing a mixed-effects regression. The parameter estimates are summarized in Table 1. In

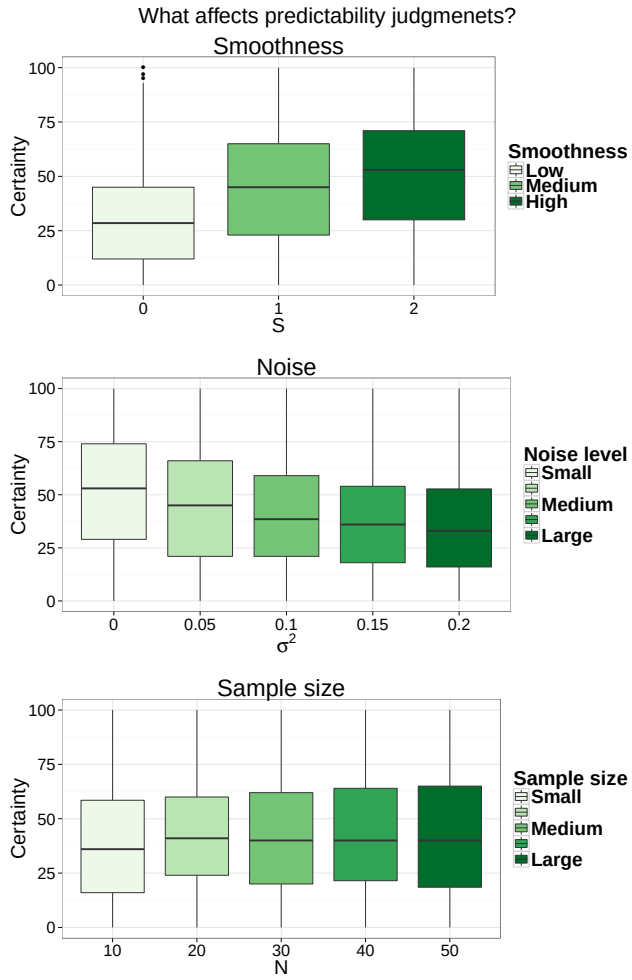

Figure 4: **Perceived predictability**.

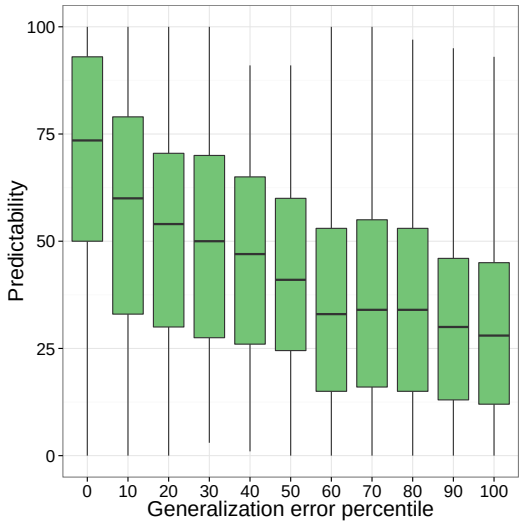

Figure 5: **Perceived predictability as a function of theoretical generalization error**.

agreement with our qualitative characterization, both smoothness and noise variance had a significant effect on predictability. The effect of sample size, on the other hand, was not significant. Nonetheless, sample size interacted significantly with both of the other variables: increasing sample size reduced the effects of smoothness and noise variance. Thus, sample size does appear to be a modulator of perceived predictability. Therefore, hypotheses 1 could be partially confirmed.

Next, we calculated the correlations between each of the different factors and the predictability judgments for each participant individually and found that the averaged correlation between smoothness and perceived predictability ($r = 0.36$, $p < 0.01$) was indeed greater than the correlation between noise and perceived predictability ($r = -0.24$, $p < 0.01$) and between sample size and perceived predictability ($r = 0.06$, $p > 0.05$). Therefore, the second hypothesis could be confirmed. Finally, we examined whether the theoretical learning curve provides an accurate quantitative model of perceived predictability. As shown in Figure 5, the theoretical generalization error is a modest but significant predictor of perceived predictability ($r = -34, p < 0.01$). Note that we used the data-independent generalization error (Eq. 7) for this analysis. However, it is more likely that participants are basing their judgments on the data-dependent generalization error (Eq. 6). Surprisingly, the data-dependent generalization error produced a slightly lower correlation ($r = -0.31$, $p < 0.01$). This suggests that we are not yet capturing some of the essential determinants of predictability perception. Therefore, the last hypothesis could only be confirmed within constraints.

Discussion

We have shown that the GP model of function learning provides a framework for understanding the perception of predictability. The model captures qualitative effects of a function's smoothness, noise variance, and sample size on the generalization error (a measure of unpredictability). The smoothness of a function exerts a stronger influence on predictability than noise or sample size, consistent with both theoretical learning curves and our experimental data. This means that a smooth but noisy function is perceived as more predictable than a complex but near-deterministic function. We also showed that the model could quantitatively capture participants' predictability judgments, although it still leaves a fair amount of variance unexplained.

One reason for the relatively low correlation between generalization error and predictability might be because our analysis assumed that the covariance function is known. If participants are using a different covariance function, this will change the form of the learning curves (although the qualitative predictions of the results re-

ported here would remain the same; Sollich, 2005). In future work, we will explore models that learn the covariance function parameters.

The answer to the question of how we perceive the predictability of a function is probably more nuanced than our account suggests. Statistically speaking, our ability to learn a function of a given complexity improves with increasing sample size, and thus for a given level of complexity there is a sample size at which we would find the function highly predictable. This means that, for example, using the same levels of complexity that we explored here but at much large samples sizes, the effect of smoothness on predictability might be relatively weaker than what we report here. At different levels of complexity and sample size, the human perception of the predictability of functions, as well as which factors drive that predictability, may vary.

While we have considered a fairly simple family of covariance functions, evidence suggests that people have richer representations; for example, a single function may be partitioned into several different sub-functions (Kalish et al., 2004). We can take this one step further and ask whether functional knowledge is compositional, building complex functions out of simpler building blocks using a 'function grammar' (Duvenaud et al., 2013). An interesting question for future research is whether people can learn complex functions more easily when they are consistent with an intuitive function grammar, similarly to how schemas facilitate the rapid acquisition of causal knowledge (Goodman et al., 2011).

Another way to explore intuitive theories of predictability is to place priors directly over the spectral density representation of a covariance function (Wilson & Adams, 2013). Because theoretical predictability can be directly related to the entropy of the spectral density (Goerg, 2013), we predict that different spectral density shapes will systematically change predictability judgments.

A different direction for future research is manipulating the way in which input points are sampled. For example, learning curves change as a function of whether inputs are sampled randomly or using directed exploration (Ritter, 2000). We intend to further validate our measure of predictability by letting participants choose between different functions and then ask them to generate predictions for newly observed points directly in a follow-up experiment. This will bring our novel approach even closer to traditional approaches of experiments on human function learning (DeLosh et al., 1997).

Unlike previous work on function learning, which has focused on interpolation and extrapolation performance, our work explored a relatively novel facet—predictability. We expect that this simple assay will provide a rich source of information about function knowledge.

Acknowledgements

ES is supported by the UK Centre for Financial Computing and Data Analytics.

References

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

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua*. Education Testing Service.

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

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, (pp. 1166–1174).

Goerg, G. (2013). Forecastable component analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, (pp. 64–72).

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110–119.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In *Advances in Neural Information Processing Systems*, (pp. 553–560).

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*, 1072–1099.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811–836.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1019–1030.

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*, 24–42.

Opper, M., & Vivarelli, F. (1999). General bounds on Bayes errors for regression with Gaussian processes. *Advances in Neural Information Processing Systems*, *11*, 302–308.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*. Springer Verlag.

Sollich, P. (2005). Can gaussian process regression be made robust against model mismatch? In *Deterministic and Statistical Methods in Machine Learning*, (pp. 199–210). Springer.

Sollich, P., & Halees, A. (2002). Learning curves for gaussian process regression: Approximations and bounds. *Neural Computation*, *14*, 1393–1428.

Williams, C. K., & Vivarelli, F. (2000). Upper and lower bounds on the learning curve for gaussian processes. *Machine Learning*, *40*, 77–102.

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, (pp. 1067–1075).