Gaussian Process Regression for Trajectory Analysis

Gregory E. Cox (grcox@indiana.edu)
George Kachergis (gkacherg@indiana.edu)
Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth St., Bloomington, IN 47405 USA

Abstract

Cognitive scientists have begun collecting the trajectories of hand movements as participants make decisions in experiments. These response trajectories offer a fine-grained glimpse into ongoing cognitive processes. For example, difficult decisions show more hesitation and deflection from the optimal path than easy decisions. However, many summary statistics used for trajectories throw away much information, or are correlated and thus partially redundant. To alleviate these issues, we introduce Gaussian process regression for the purpose of modeling trajectory data collected in psychology experiments. Gaussian processes are a well-developed statistical model that can find parametric differences in trajectories and their derivatives (e.g., velocity and acceleration) rather than a summary statistic. We show how Gaussian process regression can be implemented hierarchically across conditions and subjects, and used to model the actual shape and covariance of the trajectories. Finally, we demonstrate how to construct a generative hierarchical Bayesian model of trajectories using Gaussian processes.

Keywords: Trajectory analysis; Gaussian processes; Bayesian statistics.

Introduction

Cognitive scientists are gradually turning toward more fine-grained measures to gain more insight into the continuous nature of the cognitive processes that underly behavior. Perhaps the most widespread of these measures is eye tracking, in which we assume that where people gaze is the current focus of attention and processing. For example, when reading a syntactically ambiguous sentence, people tend to make eye movements back toward the function word or pronoun that best helps resolve the ambiguity (Frazier & Rayner, 1987). Or, when hearing continuous speech, people will tend to look more at objects whose names are consistent with a partially-heard word (e.g., people will look at either a "ball" or a "bear" if they have just heard the syllable "b"), indicating that people make continuous predictions about the content of speech based on partial information (Spivey, Grosjean, & Knoblich, 2005). Thus, a continuous measure of behavior, like eye tracking, appears to provide insight into ongoing cognitive processes.

More recently, researchers have begun to collect explicit continuous behavioral measures in the form of mouse or stylus movements (e.g., Freeman & Ambady, 2010). These may easily be used in place of any task that requires an explicit choice on the part of the participant, which includes most experimental paradigms in cognitive psychology. Rather than simply pressing a key to make their response, a participant can instead move their hand (as well as an attached mouse or stylus) toward the option of their choice before selecting (clicking) it. Similar to eye tracking, the trajectories of these continuous motor movements provide a way of measuring the ongoing cognitive processes that lead to the participant's final choice.

A major hurdle with any new measure is the need for appropriate analytical tools and statistical tests that allow researchers to draw inferences from trajectory data. Due to the richness of this data, many measures are possible and can lead to principled inferences (for an overview, see Freeman, Dale, & Farmer, 2011). When moving their hand while making a decision, people may deviate more from a straight trajectory if there is a tempting alternative, making viable such measures as maximum deviation, curvature area, and switches in direction.

In this paper, we introduce a new method for analyzing trajectory data. Our method is based on treating trajectories as a Gaussian process, for which there is much well-developed statistical theory. We begin by providing a brief overview of Gaussian process regression and show how it may be applied to motor response trajectories and—more fruitfully, we argue—their derivatives. Finally, we show how Gaussian process regression can be incorporated into a generative hierarchical Bayesian model of trajectories.

Gaussian Process Regression

Gaussian process regression (GPR) is a statistical technique with a long history in spatial statistics, and more recently in function estimation and prediction (Griffiths, Lucas, Williams, & Kalish, 2009). The interested reader is directed to the excellent text on Gaussian processes by Rasmussen and Williams (2006).

Gaussian Processes

A Gaussian process (GP) is simply a collection of random variables, all of which are jointly Gaussian distributed. What differentiates a Gaussian process from the more familiar multivariate Gaussian distribution is the fact that a Gaussian process may have an infinite index set, that is, it may specify an infinite number of jointly Gaussian variables. Thus, it is possible to define a Gaussian process over a continuous variable, like time. Just as a multivariate Gaussian distribution is defined entirely by its mean vector and covariance matrix, a Gaussian process is defined by its mean *function* $m(x)$ and covariance kernel, $k(x, x')$, where x and x' are two (possibly multidimensional) values of some predictor variable $\mathcal{X}$ (e.g., time). We denote the fact that a function $f(x)$ is a Gaussian process by

$$f(x) \sim \mathcal{GP}\left(m(x), k\left(x, x'\right)\right).$$

A Gaussian process can be considered a distribution over functions, with $m(x)$ expressing the mean value of all of these functions at x and $k(x,x')$ represented the expected covariance between the function value at x and that at x', i.e., the amount of "information" that the function $f(x')$ carries about the value at $f(x)$ (and vice versa). Thus, if we encounter data (like trajectory data) for which we do not know or cannot guess the form of the function that generated it, we can *infer* the form of this function if we assume that it is a Gaussian process. This kind of inference is termed "Gaussian process regression".

Bayesian Inference with GPs Gaussian process regression (GPR) seeks to model an unknown function $f(x)$, which is assumed *a priori* to be a Gaussian process. To do this, we need two things: a set of function observations $\mathbf{f}(\mathbf{x})$ at some known values of the predictor $\mathbf{x}$; and an expression for the covariance kernel $k(x,x')$. The data come from some experiment (e.g., a set of cursor coordinates $f(x)$). We must, however, assume a particular covariance kernel. Although many kernels are possible, for the purposes of this paper, we will confine ourselves to the squared exponential (SE) or "radial basis function" kernel:

$$k(x,x') = f \exp\left[-\frac{1}{2}\left(\frac{|x-x'|}{l}\right)^2\right]. \qquad (1)$$

The SE kernel is symmetric, is strictly positive, and most important for our purposes later, is infinitely differentiable. Notice that this kernel has two "hyperparameters": f, which scales the maximum possible covariance; and l, which functions as a length scale. Later, we will consider how the values of these hyperparameters may themselves be estimated from data, but for the moment we shall assume they are known and fixed.

Armed with a set of observations and knowledge of the covariance kernel, we now wish to perform inference on the function that is presumed to have generated the observations. In other words, we are following the logic of Bayes' rule:

$$p(\theta|\mathbf{x},\mathbf{f}(\mathbf{x})) = \frac{p(\mathbf{x},\mathbf{f}(\mathbf{x})|\theta)p(\theta)}{\int p(\mathbf{x},\mathbf{f}(\mathbf{x})|\theta)p(\theta)d\theta},$$

where θ are the parameters of the Gaussian process. Unlike in other regression settings (e.g., linear regression), where the parameters are a finite number of regression coefficients, the parameters of a Gaussian process may be infinite in number, since a GP prior allows nonzero probability to any functional form. We can, however, express our knowledge of the parameters of the GP *implicitly* via the posterior predictive distribution over *novel* function observations $\mathbf{f}(\mathbf{x}^*)$. This distribution is obtained by marginalizing over the parameters of the GP:

$$\mathbf{f}(\mathbf{x}^*)|\mathbf{x}^*,\mathbf{x},\mathbf{f}(\mathbf{x})) = \int p(\mathbf{f}(\mathbf{x}^*)|\theta)p(\theta|\mathbf{x},\mathbf{f}(\mathbf{x}))d\theta. \qquad (2)$$

This distribution captures both the residual uncertainty about the underlying function $f(x)$ *and* the knowledge gained about it from the observed data.

Posterior Predictive Distribution Computing the posterior predictive distribution begins with a prior on the mean and covariance functions of the GP, i.e., $p(\theta)$. For the moment, we shall assume that the underlying function has a constant mean of zero, with a SE covariance function (equation 1). Expressing the likelihood of the observed function values, $p(\mathbf{x},\mathbf{f}(\mathbf{x})|\theta)$, is straightforward because they are assumed to come from a Gaussian process, and hence are jointly normally distributed. The parameters of this distribution come from our prior, i.e., the prior mean of each observation is taken to be zero, and the covariance between function values is dictated by our prior covariance kernel (the SE kernel given in eq. 1). Denoting the matrix of pairwise covariances between each observed datum as $K(X,X)$, we have

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, K(X,X)).$$

Now, say we wish to express a posterior predictive distribution over function values at set of novel predictor values, denoted X^*. We can similarly compute a matrix of covariances between these points, $K(X^*,X^*)$, and between these novel points and the observed points, $K(X,X^*)$. Because both these novel points and the previously observed data values are presumed to have been generated by the same GP, they are all jointly normally distributed with mean zero and block covariance matrix:

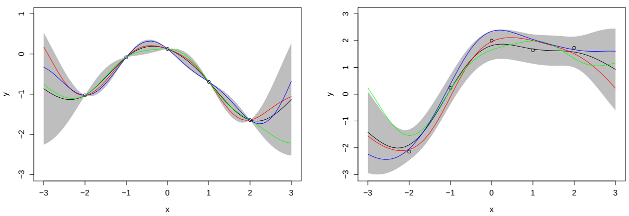
$$\begin{bmatrix}\mathbf{f}(\mathbf{x})\\ \mathbf{f}(\mathbf{x}^*)\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X^*)\\ K(X^*,X) & K(X^*,X^*)\end{bmatrix}\right).$$

We can then express the conditional posterior over $\mathbf{f}(\mathbf{x}^*)|\mathbf{f}(\mathbf{x})$ as another multivariate Gaussian distribution, using known identities regarding the Gaussian distribution:

$$\begin{aligned}\mathbf{f}(\mathbf{x}^*)|\mathbf{f}(\mathbf{x}) \sim \mathcal{N}\big(&K(X^*,X)K(X,X)^{-1}\mathbf{f}(\mathbf{x}),\\ &K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*)\big). \qquad (3)\end{aligned}$$

The posterior predictive distribution given just a few data points is shown in figure 1a. This figure also depicts three functions randomly drawn from this posterior. Note that they all pass through the observed function values (and the posterior variance at those points goes to zero). This is because we have assumed thus far that our function observations are noiseless; thus, we have absolute certainty that, whatever the true generating function is, it *must* pass through the values we have thus far observed. In addition, our (assumed) knowledge of the covariance kernel allows us to estimate the function's behavior between and, to a certain extent, beyond the observed values.

In reality, we will rarely have noiseless observations of our function of interest. Luckily, observation noise is easily incorporated into the GPR framework by adding a noise term, σ^2, to the diagonal elements of the observed covari-



(a) No observation noise.

(b) With uniform, uncorrelated Gaussian noise ($\sigma^2 = .1$).

Figure 1: Examples of GPR given a set of function observations. The open circles are the observed values. The black line is the mean of the posterior predictive distribution, while the gray region is the 95% confidence region around that mean. The three colored lines are functions randomly drawn from the posterior. Covariance was assumed to be SE with $f = 1$ and $l = 1$.

ance matrix, i.e., $K(X,X) + \sigma^2 I$. The resulting joint observed/predicted distribution becomes

$$\begin{bmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{f}(\mathbf{x}^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right)$$

and the posterior predictive distribution changes accordingly:

$$\mathbf{f}(\mathbf{x}^*)|\mathbf{f}(\mathbf{x}) \sim \mathcal{N}\Big(K(X^*,X)\left[K(X,X)+\sigma^2 I\right]^{-1}\mathbf{f}(\mathbf{x}),$$
$$K(X^*,X^*) - K(X^*,X)\left[K(X,X)+\sigma^2 I\right]^{-1}K(X,X^*)\Big). \quad (4)$$

This assumes that noise is uniformly distributed and independent between observations, but if there is correlated noise between observations, this may be incorporated directly into the covariance kernel. An example of a posterior predictive distribution with observation noise is given in Figure 1b.

GP Likelihood In order to fit a GPR model to data, we require an expression for the likelihood of a set of function observations that are assumed to come from a GP. Luckily, as is clear from above, these observations can be treated as coming from a multivariate Gaussian with mean zero and covariance matrix $K(X,X)$. Thus, the likelihood is merely the multivariate Gaussian likelihood:

$$p(\mathbf{f}(\mathbf{x})) = (2\pi)^{-\frac{n}{2}}\,|K(X,X)|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\mathbf{f}(\mathbf{x})^T K(X,X)^{-1}\mathbf{f}(\mathbf{x})\right] \quad (5)$$

where n is the number of observed data points and $K(X,X)$ may be replaced by $K(X,X) + \sigma^2 I$ if observation noise is assumed.

Multiple Observed Functions If multiple functions are observed simultaneously, e.g., the x and y coordinates of a cursor on a screen, they can each be treated as *a priori* independent Gaussian processes and the above reasoning applied to each individually.

Derivatives of Gaussian Processes

Because differentiation is a linear operation, the derivative of a GP is itself a GP. Function derivatives are useful in the event that we actually have observations of the derivative (as in Solak, Murray-Smith, Leithead, Leith, & Rasmussen, 2003). However, we also argue that the derivatives of a continuous response like a mouse movement are more informative about the underlying cognitive process that generates them. For example, the acceleration is critical for finding inflection points, which could indicate that the participant is considering changing his or her mind, or that they have just incorporated new information into their decision process.

In the cases we consider below, we have direct observations only of position information, not of its derivatives (e.g., velocity and acceleration). To compute a posterior predictive distribution over function derivatives, we need only compute the covariances between each function observation and the its derivatives at the points at which we are seeking predictions. This, in turn, requires expressions for the covariances between function values and derivatives, which are given for the SE covariance kernel below:

$$\frac{\partial}{\partial x}k(x,x') = -\frac{k(x,x')}{l^2}\left(x-x'\right) \quad (6)$$

$$\frac{\partial^2}{\partial x^2}k(x,x') = \frac{k(x,x')}{l^2}\left[\left(\frac{x-x'}{l}\right)^2 - 1\right] \quad (7)$$

$$\frac{\partial^2}{\partial x \partial x'}k(x,x') = \frac{k(x,x')}{l^2}\left[\left(\frac{x-x'}{l}\right)^2 + 1\right] \quad (8)$$

$$\frac{\partial^4}{\partial x^2 \partial x'^2}k(x,x') = -\frac{k(x,x')}{l^4}\left[\left(\frac{x-x'}{l}\right)^2\left(3 - \left(\frac{x-x'}{l}\right)^2\right) + 3\left(\frac{x-x'}{l}\right)^2 - 3\right]. \quad (9)$$

We can then compute a posterior predictive distribution over any desired derivative, given only raw function observations, by constructing the covariance matrices $K(X^*,X)$ and $K(X^*,X^*)$ from equation using the appropriate partial derivative above, rather than the original SE kernel $k(x,x')$. For example, to compute the posterior predictive distribution for the velocity, $\dot{\mathbf{f}}^*(\mathbf{x}^*)|\mathbf{f}(\mathbf{x})$, compute $K(X^*,X)$ using equation 6 for each pair of predicted and observed x values and $K(X^*,X^*)$ using equation 7 for each pair of predicted x values.

Applications of GPR to Trajectory Analysis

In this section, we provide several examples of applications of GPR to trajectory analysis. In so doing, we introduce several extensions to the GPR modeling framework that place it in the realm of hierarchical generative models which can enable principled Bayesian inferences regarding the cognitive processes that underly observed motion trajectories.

Estimating Hyperparameters

Although the posterior distribution in GPR is easily expressed analytically given knowledge of the covariance kernel and its

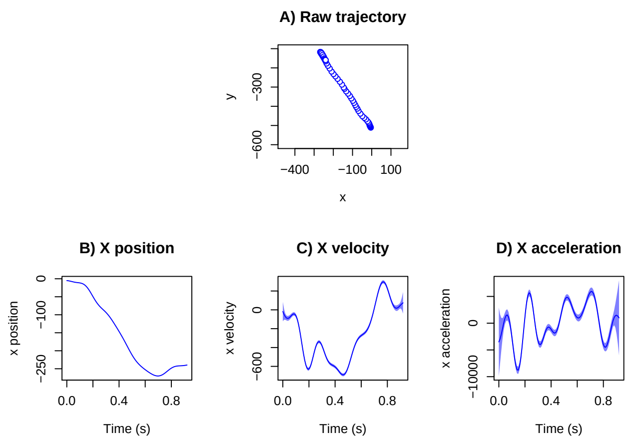


Figure 2: Example of GPR on a single two-dimensional trajectory. In B, C, and D, the light blue region depicts a 95% credible region about the mean posterior predictions.

hyperparameters, we are left with having to estimate f, l, and σ^2 (if we assume there is noise in the observations). In particular, we would like to be able to express our beliefs over these hyperparameters in the form of a posterior distribution. Unfortunately, this posterior will not in general be expressible analytically. Thus, we must turn to Monte Carlo methods to estimate the posterior over these parameters.

A Single Trial Let us assume we have a single mouse trajectory in two dimensions, as shown in Figure 2A. This trajectory is a single trial from the experiment reported by Spivey et al. (2005). On each trial of this experiment, participants saw two objects, the names of which either had similar phonological onsets (i.e., were members of the same "cohort") or had phonologically unrelated names (the control condition). An audio recording instructed the participant to move their mouse cursor from a box in the lower center of the screen and click on one of the two objects (thus ending the trial).

The single trajectory consists of a series of (t, x, y) triples, with x and y coordinates and the times t at which they were observed. We treat the times $\mathbf{t}$ as a univariate predictor (i.e., in the role of x in the previous section) and x and y as conditionally independent Gaussian processes operating on t, with zero mean and SE covariance kernel (i.e., in the role of $f(x)$ in the previous section). There are three hyperparameters that must be estimated: the parameters of the covariance kernel, f and l (see equation 1), and a noise term, σ^2, which is assumed to apply to measurements in both the x and y directions (isotropic noise is assumed here merely for simplicity). We choose very vague priors on each of these hyperparameters, such that they are informed almost entirely by the data, rather than our priors (although these priors could be informed by knowledge, e.g., of the accuracy of mouse position measurements). We assign a Gamma prior to f and l with shape and scale parameters set to 0.001 and an inverse-Gamma prior to σ^2 (also with shape and scale parameters of 0.001). The likelihood of the observed trajectory, conditional on particular values of the hyperparameters, is then given by equation 5. This model was implemented in JAGS (Plummer, 2011), drawing 1000 samples from the joint posterior over hyperparameters after 1000 steps of "burn-in".

The estimated posterior mean of each hyperparameter is $\bar{f} = 14760$, $\bar{l} = 0.1146$, and $\bar{\sigma}^2 = 0.9471$. The bottom three graphs of Figure 2 (B, C, and D) show the mean and 95% credible region of the posterior predictive distribution for the x coordinate (as well as its velocity and acceleration), marginalized over the samples of the hyperparameters.

Multiple Trials While this simple example illustrates how GPR can be applied to a single trajectory, we usually have several trials per participant per condition. In this case, we have multiple sets of triples, $\{(\mathbf{t_1}, \mathbf{x_1}, \mathbf{y_1}), (\mathbf{t_2}, \mathbf{x_2}, \mathbf{y_2}), \ldots, (\mathbf{t_n}, \mathbf{x_n}, \mathbf{y_n})\}$, and we can treat them all as having been generated by the same underlying GP. In other words, even if two observations (x_1, y_1) and (x_2, y_2) were from different trials, we can still compute their covariance $k(t_1, t_2)$ as a function of the times t_1 and t_2 at which they were observed, as if they were part of the same trial (and thus they also share hyperparameters). Collecting the observed function values $\mathbf{x_i}$ and $\mathbf{y_i}$ (where i indexes the trial), we can write

$$\begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_n} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{x_1}, \mathbf{x_1}) & K(\mathbf{x_1}, \mathbf{x_2}) & \cdots & K(\mathbf{x_1}, \mathbf{x_n}) \\ K(\mathbf{x_2}, \mathbf{x_1}) & K(\mathbf{x_2}, \mathbf{x_2}) & \cdots & K(\mathbf{x_2}, \mathbf{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x_n}, \mathbf{x_1}) & K(\mathbf{x_n}, \mathbf{x_2}) & \cdots & K(\mathbf{x_n}, \mathbf{x_n}) \end{bmatrix}\right)$$

where $K(\mathbf{x_i}, \mathbf{x_j})$ denotes the covariance matrix between each sample in trial i and trial j (and a similar multivariate Gaussian likelihood is defined for $\mathbf{y}$).

The assumption leading to the above likelihood is only valid if we assume that trajectories generated by the same participant in the same condition in fact represent samples from the same underlying process. If we assume that different trials from the same participant may be come from different processes that nonetheless share some underlying characteristics, the hierarchical extension of GPR that we introduce in the next section may be employed instead.

Hierarchical GPR

Having shown how GPR can be applied to single trajectories and to multiple trajectories that may be assumed to share the same hyperparameters (i.e., to have been generated by the same underlying GP), we now turn to the case of multiple conditions and multiple participants per condition.

Multiple Conditions When there are multiple conditions in an experiment, we assume that a trajectory produced in one condition is conditionally independent of a trajectory produced in another condition, that is, that the trajectories are generated by different GP's that nonetheless share hyperparameters. The rationale for sharing hyperparameters across conditions is simple: measurement noise (the σ^2 parameters,

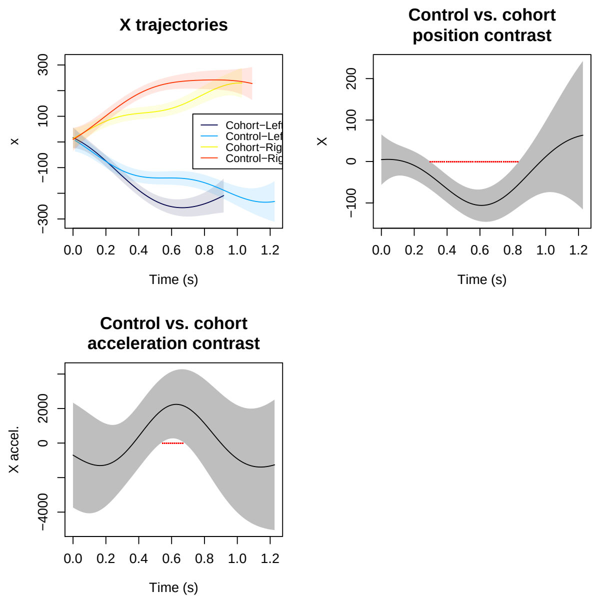


Figure 3: Posterior predictive distributions for trajectories inferred from a single subject from Spivey et al. (2005). The upper right plot shows the contrast computed between the x-positions in two cohort and two control conditions, while the lower left plot depicts the same contrast with the x-accelerations. Solid lines show the posterior predictive mean while the colored regions depict 95% credible regions around the corresponding mean.

one for each observed function) should depend only on the apparatus (e.g., the mouse or stylus). The hyperparameters of the covariance kernel, meanwhile, may be interpreted to reflect properties of the motor system of the participant, which are, of course, shared across conditions: f reflects the degree of "hysteresis", or the tendency for the participant to produce trajectories with points that lie near one another, while l is indicative of the typical size of deviations from a straight line[1].

We can again express the conditional likelihood of a set of function observations as a zero-mean multivariate Gaussian. Similar to the multiple-trial situation above, we can denote the observed trajectory points in condition j by $\{\mathbf{x_i}\}_j$ and the covariance between each observation in condition j as $K(\{\mathbf{x_i}\}_j, \{\mathbf{x_i}\}_j)$. Then, we construct a block covariance matrix for the likelihood that reflects our assumptions about conditional independence between conditions:

$$\begin{bmatrix} \{\mathbf{x_i}\}_1 \\ \{\mathbf{x_i}\}_2 \\ \vdots \\ \{\mathbf{x_i}\}_n \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\{\mathbf{x_i}\}_1, \{\mathbf{x_i}\}_1) & 0 & \cdots & 0 \\ 0 & K(\{\mathbf{x_i}\}_2, \{\mathbf{x_i}\}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K(\{\mathbf{x_i}\}_n, \{\mathbf{x_i}\}_n) \end{bmatrix}\right)$$

And, again, we can follow the same logic to construct a similar likelihood for y or any other observed component of the trajectory.

To assess differences in trajectories between each condition, we can compute functional "contrasts" by taking the difference of the posterior predictive distributions for two conditions. This is done for one subject's data from Spivey et al. (2005) and shown in Figure 3. In this case, the contrast is between two pairs of conditions: the two cohort conditions (left and right) and the two contrast conditions. Zero lies outside the 95% credible region of the position contrast function between roughly .30 and .83 seconds, indicating that the trajectories produced by this subject to cohort and control stimuli are credibly ("significantly") divergent over this region. This divergence results from the additional complexity of the cohort trajectories, which is shown by the acceleration contrast: The cohort trajectories include an additional "nudge" between .55 and .67 seconds after stimulus onset, as the subject reconsiders what he or she has heard.

This example illustrates two useful features of GP's as trajectory models: First, when analyzing contrasts, they do not risk inflating the probability of false alarms due to comparisons at multiple time-points. Because a GP represents a distribution over *functions*, there is only one comparison actually taking place. Second, because GP's allow one to compute the higher derivatives of a trajectory, they afford greater insight into the functional behavior that gives rise to observed differences between trajectories, leading to potentially useful insights into the cognitive processes that generate them.

Multiple Participants We further expand the scope of the analysis by allowing for multiple participants, each of whom contributes data in multiple conditions, perhaps in many trials. Researchers typically obtain trajectory measurements from multiple participants in the same condition in order to better estimate a general property that is presumed to hold across the population. In a memory experiment, this general property might be the probability of correctly recognizing a previously studied item. There may be great variability between participants in their ability to recognize the item, but each observation is presumed to be a sample from a general group tendency.

In the case of trajectory analysis, we similarly assume that each participant produces a trajectory (or trajectories) that are samples from a distribution of possible trajectories. A GP expresses just such a distribution. Hence, we assume that there is a group-level GP for each condition, the covariance kernel of which has its own hyperparameters f_G and l_G. This group level GP captures the covariance between different trials generated by different participants in the same condition. Meanwhile, the covariance between different trials generated by the *same* subject have their own covariance structure that is added to the group-level covariance. For example, if x_1 and x_2 are two data points observed in different conditions, their covariance $k(x_1, x_2) = 0$, as before. If, however, x_1 and x_3 come from the same condition, but different subjects, their covariance will be a function of the group-level covariance, parameterized by f_G and l_G, denoted $k_G(x_1, x_3)$. Finally, if x_1 and x_4 are two data points generated by the same subject (subject s) in the same condition, their covariance will be the group covariance plus the covariance resulting from individual variation around the group trajectory, i.e., $k_G(x_1, x_4) + k_s(x_1, x_4)$, where $k_s(\cdot, \cdot)$ is a covariance kernel parameterized by subject-

[1] Of course, other choices of covariance kernel would have their own parameters which would have their own characteristic interpretations.

specific parameters f_s and l_s. As before, we can construct from these terms a covariance matrix for the entire dataset.

To perform inference in this case requires placing priors on both the group-level hyperparameters and the subject-level hyperparameters. When using vague priors, as we have thus far, it is often advisable to make use of hyperpriors. Thus, we let $\mu_f \sim \text{Gamma}(0.001, 0.001)$ and $\sigma_f^2 \sim \text{Inverse Gamma}(0.001, 0.001)$ be top-level priors on the mean and variance of the distribution of f_s values per subject. Then, by moment matching, we draw each $f_s \sim \text{Gamma}\left(\frac{\mu_f^2}{\sigma_f^2}, \frac{\mu_f}{\sigma_f^2}\right)$. We do the same for each l_s (i.e., place a hyperprior on the mean and variance). By using hyperpriors in this way, we obtain "shrinkage" of the estimates of the subject-specific parameters, such that they can mutually inform one another.

Generative GPR

Thus far, we have employed GPR solely in the way it was originally intended: as a nonparametric approach to function approximation—as a purely *descriptive* model. However, we can use GPR as a *generative* model in the following way: Say that we expect all trajectories in a particular condition to possess characteristic *landmarks*. These landmarks may be actual positions, or they may be particular values of one of the derivatives of the trajectory. For example, an inflection point—a point where the acceleration of the trajectory in a particular direction reverses—may have a special cognitive interpretation. In our ongoing example from Spivey et al. (2005), such a point may reflect the instant at which the word in the cohort condition has been completely processed, and the participant moves his or her cursor away from the distractor and toward the named object.

To implement this idea in a Bayesian fashion, we place a prior on the number of inflection points in the group-level trajectory. In principle, this number could be infinite, but in practice we assume this is a multinomial draw between 1 and 8 (the maximum allowed number of inflection points may, of course, vary depending on application). This multinomial is, in turn, parameterized by a draw from a Dirichlet distribution, which itself reflects a prior on the overall probability that the trajectory has a certain number of inflection points (from 1 to 8). Finally, for a given number of inflection points, the points themselves are presumed to be *a priori* uniformly distributed in time across the range of data points.

We can make use of the same formalism we have previously used to obtain the posterior predictive distribution for GPR to compute the likelihood, conditional on a certain set of sampled inflection points $i_1, i_2, \ldots, i_n$. This involves computing the covariance matrix $K(i,i)$ between the inflection points using the kernel in equation 9 and between the inflection points and observed values $K(X,i)$ via equation 8. The conditional covariance of the data is then $K(X,X) - K(X,i)K(i,i)^{-1}K(i,X)$. The posterior predictive distribution, along with a sample of inferred inflection points, is shown in Figure 4. Notice that only the cohort conditions have inflections in the central region, reflecting the greater complexity of both the trajectories and their underlying cognitive processes.

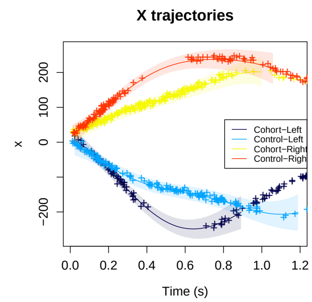


Figure 4: Posterior predictive distributions with a sample of the inflection points (+) inferred by the generative model.

Discussion

We have presented a general statistical method for modeling trajectories, and shown how it can be used to capture effects at multiple levels. A main advantage of Gaussian process regression over the various summary statistics used previously (e.g., maximum deviation) is that less information is thrown away: looking at the posterior density shows a normatively correct summary of the data, given the general assumptions made by the model. GPR balances functional complexity with capturing the underlying data, and is thus both more general and more principled than other forms of regression. Hierarchical GPR may be used to distinguish individual, group, and condition differences.

By accurately tracing and modeling the movement of the body, we can find evidence of ongoing cognitive processes, and literally see the shape of their influence. Many have wondered when psychology will reach paradigmatic maturity–like physics. Trajectories, tracing movement through space over time, are a fundamental property that all organisms and matter create.

References

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*(5), 505–526.

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*(0).

Griffiths, T. L., Lucas, C. G., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems*, *21*.

Plummer, M. (2011). *JAGS: Just another gibbs sampler.* Available from `http://mcmc-jags.sourceforge.net/`

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: The MIT Press.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., & Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In S. T. Becker & K. Obermeyer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 1033–1040). MIT Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398.