Jodie Simkoff
Gaussian Process HW
November 8th, 2017

Introduction

Gaussian processes provide a non-parametric representation of data generation processes, wherein the data are assumed to be sampled from a multivariate Gaussian distribution. Since Gaussian processes are fully defined by their first and second moments, only the mean and covariance are needed to fully define the behavior of a given GP. The kernel function specifies how the covariance of the GP is generated, and acts as a prior for the GP. These kernel functions include hyperparameters which are optimized to best fit the data.

In this experiment, we have motion data corresponding to the position of a certain part of the body in a given coordinate, over time. The data was collected as the subjects traced curves, and the goal is to see if we can represent their motion as a Gaussian process. Since this data set is very complex, we focus on a single coordinate for a single marker: we choose marker 4, which is located on the subjects' heads, and we consider motion in the x direction.

Method Description

We considered a squared exponential kernel function:

$$k(t,t^*) = \exp(\sigma_f)\exp(-0.5\exp(\sigma_l)\,|t-t^*|^2) + \exp(\sigma_n)\,\boldsymbol{I}$$

Where t and t^* are time vectors, $\sigma_f, \sigma_l, \sigma_n$ are hyperparameters representing, respectively, covariance of the processes at a given input, dependence of outputs on the separation in their respective inputs, and assumed noise in the data. Thus, for given time vectors of length n, the kernel function produces an nxn covariance matrix.

In GP regression, it is assumed that the data is a sample from a multivariate Gaussian. So, the probability of a prediction of data at some input can be conditioned on the distribution of the observations. The conditional probability can be stated as

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}$$

$$p(y^*|y) \sim N(K^*K^{-1}y, K_{**} - K_*K^{-1}K_*^T)$$

However, before predictions can be made, the hyperparameters need to be learned. This can be done by using the log-likelihood function of the Gaussian process model, and maximizing with respect to the hyperparameters given some training data set (y,t).

The log likelihood function is:

$$\log P(y|f,\theta) = 0.5 f^T K^{-1} f - 0.5\log|K| - constant$$

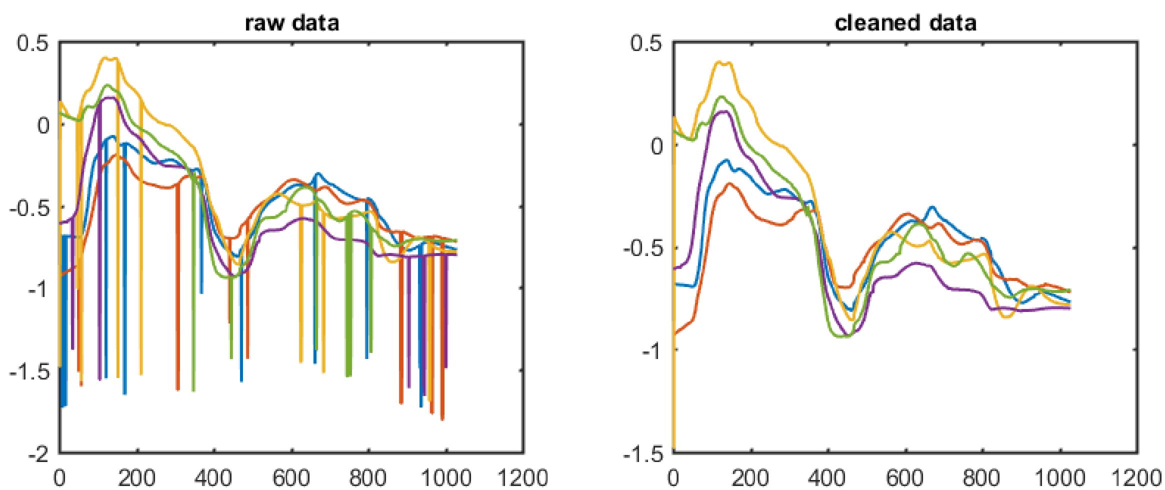
And the gradient with respect to theta is

$$\frac{d}{d\theta}\log P(y|f,\theta) = -0.5trace\left(K^{-1}\frac{dK}{d\theta}\right) + 0.5f^T K^{-1}\frac{dK}{d\theta}K^{-1}f$$

Note that this also requires the gradients of the kernel function with respect to hyperparameters. In this assignment, we use the function gprfit() in MATLAB® to find the optimal hyperparameters given the above kernel function and the training data. As a result, the log-likelihood values during the gradient descent iterations are not plotted.

Data cleaning

Since the data sets contain some faulty measurements, denoted by negative values of the c data field, the data were first smoothed by replacing faulty points with an estimated obtained using linear interpolation from the nearest non-faulty data points. A raw and smoothed data set are example shown below:



Overview

To investigate the head (marker 4) x-movement data as a Gaussian process, we do the following:

1. Fit single subject, single trial as a Gaussian process; use to predict data at other time points and establish confidence intervals.
2. Fit single subject, multiple trials and find the resulting GP. Compare to the single subject results.
3. Fit multiple subjects and multiple trials to find the resulting GP. Compare to single subject results.
4. Consider the effect of splitting a single trial into multiple sections – do the optimal hyperparameters change across the time horizon?

Subject SG: Trial 1

The hyperparameters for the Gaussian process were fit using gprfit() in MATLAB®. The training data points were obtained by taking every 40th measurement, resulting in 26 total training points.

Figure 1 shows the fit to the training data at top left, and the interpolated predictions (at **test points**) using bad initial hyperparameters (top right) and using the optimized hyperparameters (bottom left). The test points are evenly spaced between the training points. A zoomed in view, i.e. over a smaller range of test points, is shown in the bottom right plot. The optimal hyperparameters are as follows:

Subject SG: Trial 1	
Hyperparameter	**Value**
$\exp(\sigma_f)$	0.10806
$\exp(\sigma_l)$	0.19171
$\exp(\sigma_n)$	0.0014056

Table 1. SG Trial 1 hyperparameters

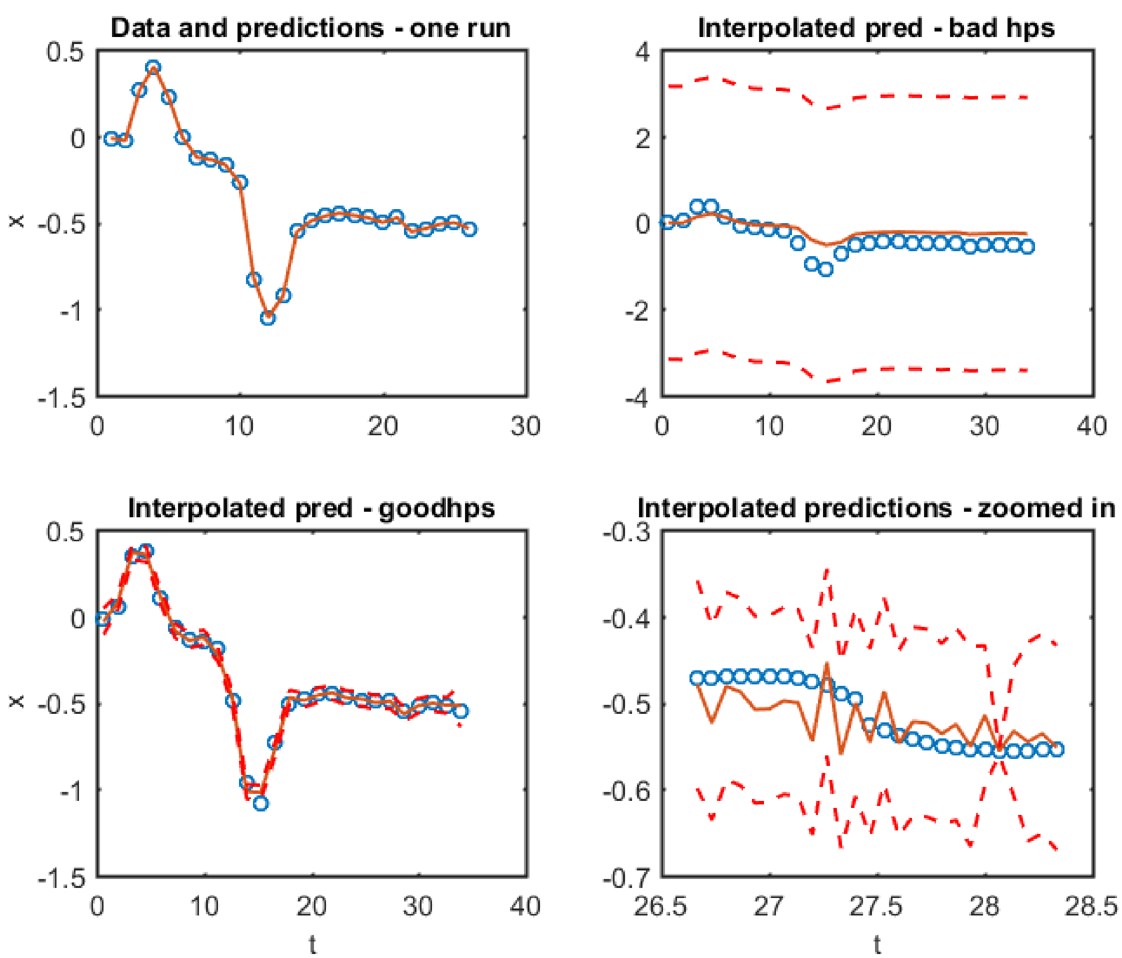


Figure 1. Blue dots are data points, orange is predicted by GP, red dashed are 90% confidence intervals

Subject SG: All trials

Next, all trials for subject SG were used to fit a GP, using the mean values.

The results are shown in Figure 2 and Table 2. Note that the hyperparameters are different between the single trial and averaged result, especially in the length and noise parameters. Still, the results are same order of magnitude and considerably similar given that in each case the gradient descent was initialized with values very far from these optimal values.

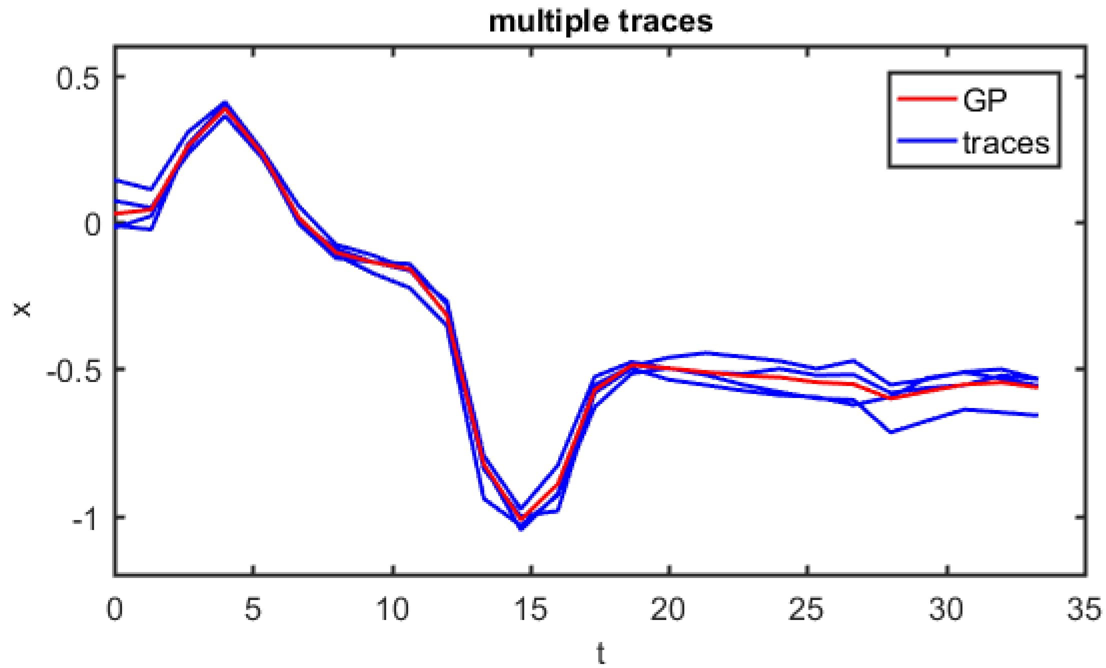


Figure 2. Average GP obtained from all traces for subject SG, marker 4-x

	Subject SG: Trial 1	**Subject SG: All trials**
Hyperparameter	**Value**	**Value**
$\exp(\sigma_f)$	0.10806	0.11380
$\exp(\sigma_l)$	0.19171	0.16286
$\exp(\sigma_n)$	0.0014056	0.0010719

Table 2. SG hyperparameters

Three subjects: All traces

Next, the same was done for 15 trials: all trials' 4-x data from subjects SG, MC, and CJ.

	Subject SG: All trials	Subject MC: All trials	Subject CJ: All trials	Subject SG, MC and CJ: All trials
Hyperparameter	**Value**	**Value**	**Value**	**Value**
$\exp(\sigma_f)$	0.11380	0.09403	0.067553	0.076950
$\exp(\sigma_l)$	0.16286	0.19021	0.15865	0.17655
$\exp(\sigma_n)$	0.0010719	0.00067648	0.00048620	0.00047045

Table 3. Multiples trials, multiple subjects

The optimal hyperparameters for the average processes per subject and for the three subjects taken together are shown in Table 3, and the results are plotted in Figure 3. Overall, the traces are remarkably similar across subjects and trials, and so a Gaussian process with the optimal hyperparameters found via this regression is likely to be extendable to predicting results from other trials as well.

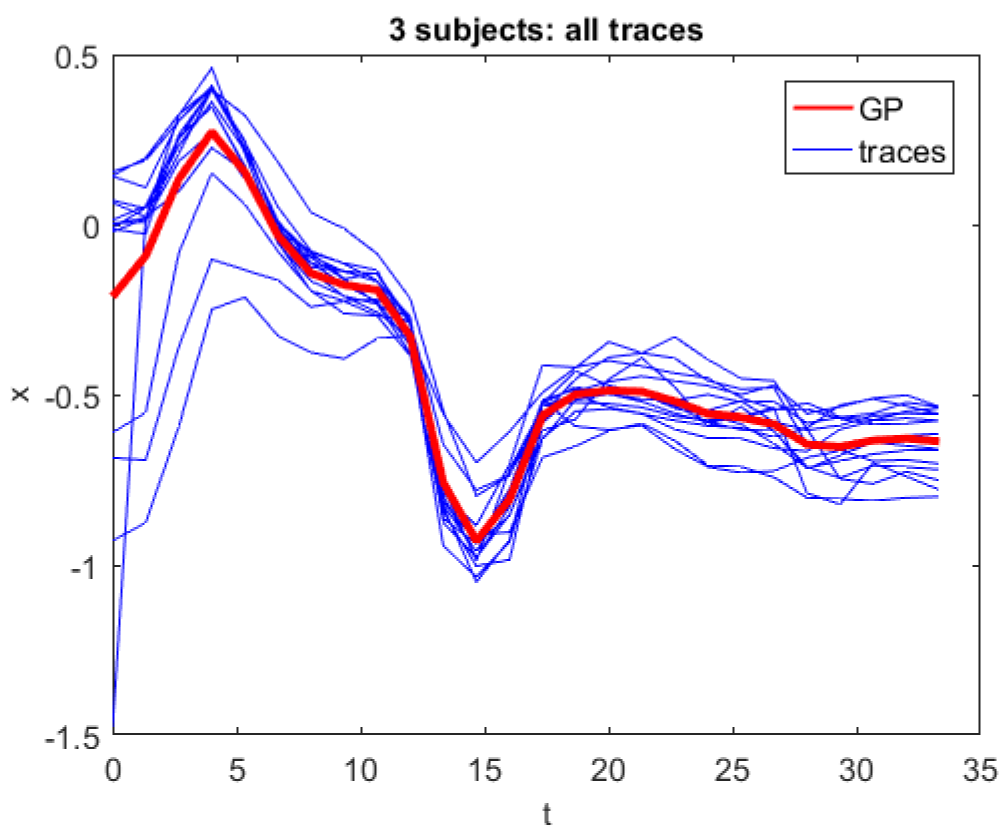


Figure 3. A GP fit based on all trials from subjects SG, CJ, MC

Time-varying hyperparameters

Next, the possibility of time-varying optimal hyperparamters was explored. Visually, the x-coordinate in the first half of the data changes quite a lot, while in the second half, the value is pretty constant. So, the data was split into two sections, and every 20th sample point was used to train a GP in the respective section. The results are plotted in Figure 4 and the values of the hyperparameters are reported in Table 4. The results are quite different in each half of the data. In the second half, the value of $\exp(\sigma_l)$ is considerably higher (0.97) than in the first half (0.21). Clearly the correlation of data points across time is higher in the second half as compared to the first, which can be verified in the trace plots. As such, this segmentation strategy might be a reasonable approach for representing the 4-x positions as a series of Gaussian processes. The log-likelihood values are also presented in Table 4 and show that the data segmentation significantly increases this measure of the fit of the data to the model.

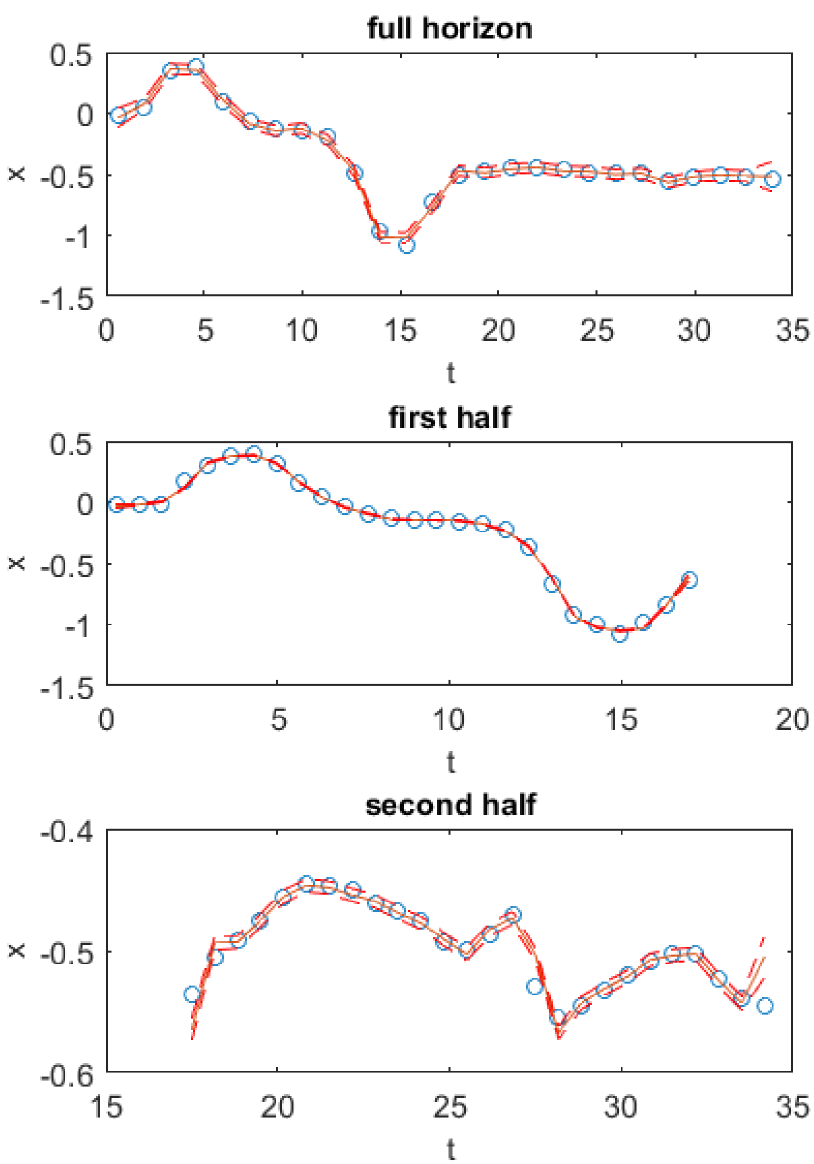


Figure 4. Time-varying hyperparameters: plots of interpolated predictions. Blue dots are data points, orange is predicted by GP, red dashed are 90% confidence interva

	Subject SG: trial 1, full	Subject SG: trial 1, first half	Subject SG, trial 1, second half
Hyperparameter	**Value**	**Value**	**Value**
$\exp(\sigma_f)$	0.10806	0.16105	0.0014976
$\exp(\sigma_l)$	0.19171	0.20824	0.97181
$\exp(\sigma_n)$	0.0014056	0.000091207	0.000013706
Log-likelihood	14.08	29.47	63.15

Table 4. Time-varying hyperparameters

Finally, a the predictive ability was tested by considering a small segment of time in the second half of the horizon and comparing the predictions from the full horizon and second half GPs, along with their confidence intevals. This is shown in Figure 5. The second half fit for the hyperparameters produces a much better representation of the interpolated points. This can be attributed to the fact that the noise term is lower while the length-variance term is higher in the second half than for the full horizon fit.

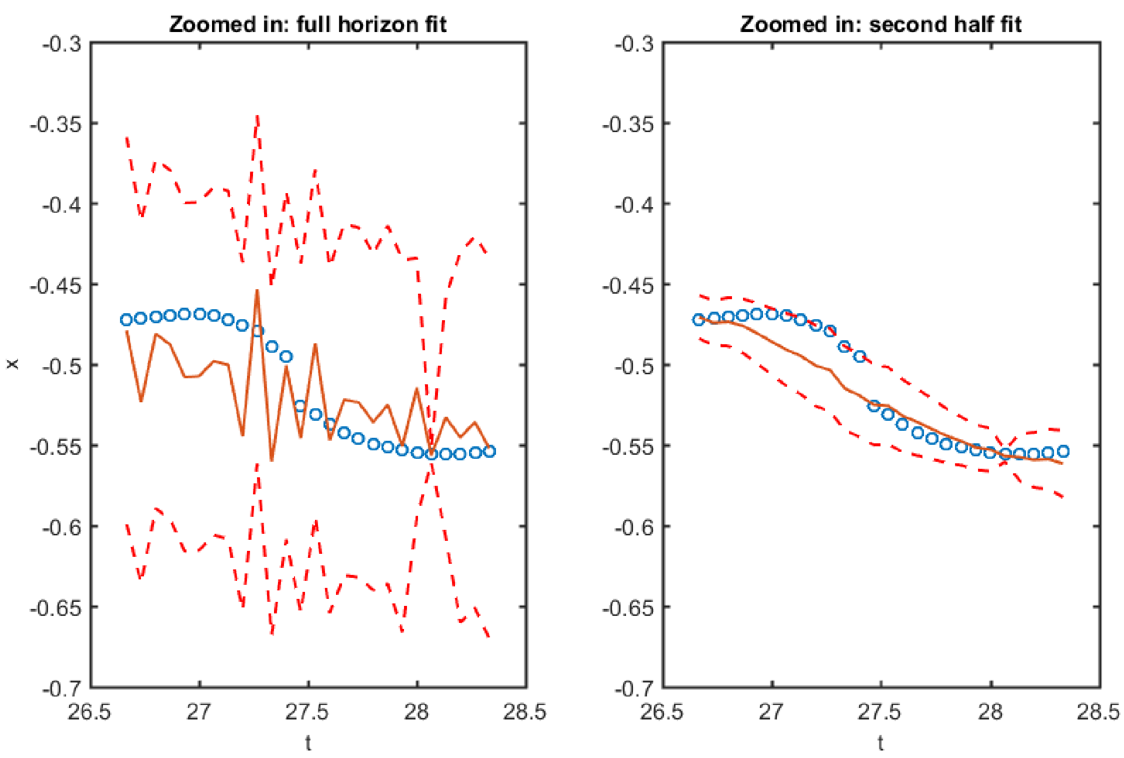


Figure 5. Comparison of predictive performance of full-horizon and second-half Gaussian processes. Blue dots are data points, orange is predicted by GP, red dashed are 90% confidence intervals

Conclusions

In this work, we have use Gaussian process regression to represent the time-varying x-coordinate of a sensor located on a subject's head during a tracing experiment. Via this regression, the optimal hyperparameters of the GP were determined and were used to predict the data at test points by maximizing the probability of the test point given the distribution fit to the training points. The traces produced by the sensor were shown to be quite similar across trials and subjects and so a representative Gaussian process was developed using multiple trials. It was also shown that for this particular example, the observed behavior trends change with time, and so the use of time-varying hyperparameters – essentially representation by multiple Gaussian processes – improves the model fit.