UNIVERSITY OF CALIFORNIA
SANTA CRUZ

GAUSSIAN PROCESS KERNEL SELECTION FOR PERFORMANCE PREDICTION BASED ON PHYSIOLOGICAL DATA

A thesis submitted in partial satisfaction of the
requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL & COMPUTER ENGINEERING

by

Jonathan M. Wagstaff

June 2022

The Thesis of Jonathan M. Wagstaff
is approved:

Professor Steve McGuire, Chair

Professor Sri Kurniawan

Professor Dejan Milutinovic

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Gaussian Process Kernel Selection for Performance Prediction Based on Physiological Data

by

Jonathan M. Wagstaff

The adoption of Human-Robot (HR) teaming continues to increase within many high-risk fields (e.g., space exploration, medical treatment). This work solves an important problem in HR teaming by learning to interpret Mental Workload (MW) from human passive biosignals in the context of human performance. Using a previously designed experiment, we analyze GP kernels for human performance estimation. GP models help limit designer bias and provide prediction confidence intervals. This analysis offers the following contributions to this field: a heuristic to help understand which biosignals are informative, an evaluation of data, and a comparison of GP kernels. The experiment showed that smooth kernels yield lower performance prediction Root Mean Squared Error (RMSE) and Standard Error of the Mean (SEM) for each performance metric considered. As a result of this work, performance prediction model designers will have a guide for improving HR systems, passive MW monitoring, and performance estimation of HR teaming.

Chapter 1

Introduction

MW models provide robotic systems informative data to increase performance within HR teams in a variety of high-risk missions. For example, scenarios involving rescue missions [27], driving [7], aviation (both manned [2, 33] and un-manned [41] aircraft), and surgery [45] have shown interest in MW evaluation. Studies measure MW, the ratio of demand and allocated resources [9], through periodic surveys or continuous physiological monitoring while a subject completes various tasks [6]. While surveys produce a subjective validation for MW models, physiological responses provide continuous, objective data. Many studies use physiological measurements, or biosignals, to estimate MW which then correlates to task performance [18, 43]. Poor MW modeling can lead to increases in cognitive demand and negative emotional state [15], decreases in safety and performance [43], delayed informational processing [32], and an inability to resume a task [13]. Adoption of MW models will require operator confidence in model predictions, which interpretable models can provide. In order to mitigate risks and

improve HR teaming performance, designers need to optimize models for physiological data while maintaining prediction interpretability.

Useful MW models require data that reflect realistic responses to task demands from a variety of biosignal sources. Physiological responses to MW manifest in several ways, including changes in brain, cardiac, respiration, skin, and ocular activity [6]. Research shows that creating multimodal models from different biosignals may yield better MW estimates [3, 10]. Collection of realistic physiological data requires test subjects to experience authentic task demands. Virtual Reality (VR) has become an immersive, experiential environment to induce genuine responses [23], especially for high operational cost and high-risk tasks (e.g. fighter pilot missions). VR simulations help immerse subjects into the environment and conveniently permit subjects to wear a variety of passive biosensors.

GPs, a probabilistic machine learning (ML) technique, estimate functions with confidence intervals meeting the criteria for a model with interpretable predictions. GPs have also proven effective for biosignals [5]. While deterministic models require a large amount of training data, GPs do not require any training data making them good candidates for online modeling. As a GP designer, the primary decision in creating a GP model is the choice of kernel, which defines similarity between data points [11, 29]. We are unaware of any study that has investigated which kernel yields the best performance predictions based on physiological data.

This study explores the optimization of online GP performance predictions for different kernels based on a VR experiment data set [42]. The main contributions of

this work are a heuristic for determining which biosignals supply useful insight to MW and a comparison of GP performance based on kernel selection. Chapter 2 details the mechanics of measuring mental workload and introduces GP kernels and the VR data set used in this study. Chapter 3 clarifies the scope of this study. Chapter 4 explains the procedure for data processing, GP model development, and kernel evaluation. Chapter 5 examines the kernel evaluations and implications, with details of this work's greater impact considered in Chapter 6.

Chapter 2

Previous Work

The subject matter for this research stems from two primary fields: cognitive state estimation (CSE) from physiological measurements and GP kernel selection. MW, synonymous with Cognitive State or Cognitive Load in this paper, has been defined as the ratio of demand to allocated resources [9]. In particular, MW considers an individual's reaction to the demands of a task. GPs are a ML technique that estimates functions to fit a data set. This chapter reviews and discusses previous studies that incorporated these topics as well as the origin of the data used in this study.

2.1 Measuring Mental Workload

Physiological signals from cardiac, blood pressure, skin, respiratory, ocular, and brain measurements give quantitative insight into an individual's MW [6]. For most raw physiological data, features are extracted to interpret and detect physiological changes. For example, breathing rate is extracted from a respiratory waveform. The

most common measurement in CSE is electrocardiac activity measured by Electrocardiogram (ECG) [6]. Multimodal models based on more than one biosignal lead to better estimates [3, 10]. The physiological measurements discussed in this section are limited to the set of features from the data set used in this study.

2.1.1 Cardiac Measurements

Changes in cardiovascular activity provide indirect insight into MW [6]. ECG, the most common measurement of heart activity in CSE, records the electrical activity of the heart. Multiple electrodes rest on the body and measure the electrical potential between themselves. The ECG waveform can be decomposed into a chain of smaller waves (see Figure 2.1). Previous extensive studies have composed a standard list of useful features from the frequency, magnitude, and variation of the ECG waveform [35].

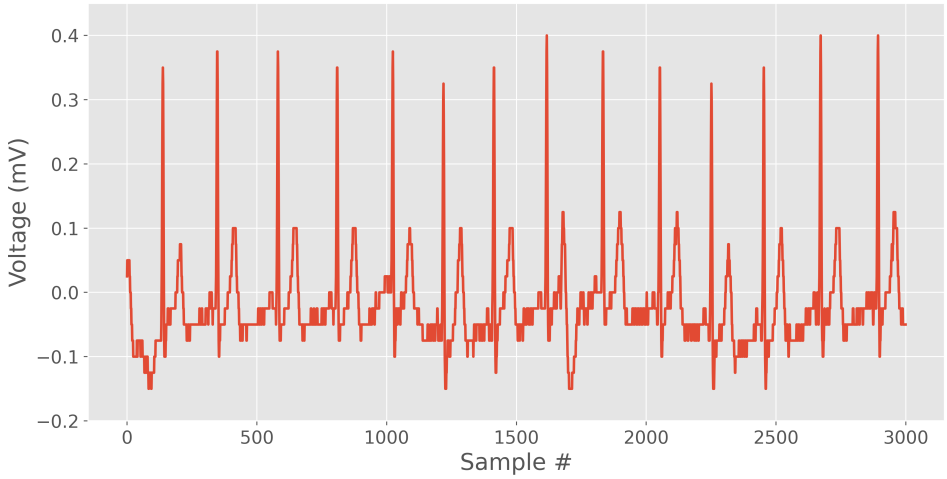


Figure 2.1: Raw ECG waveform. Most features extracted from the ECG waveform consider the timing and frequency of wave peaks.

ECG feature extraction produces measures in both the time and frequency domains. Temporally, ECG measures (Heart Rate and Heart Rate Variability (HRV)) have demonstrated relationships to MW [6, 30, 39]. In the frequency domain, power bands increase or decrease as a function of task load [6]. Notably, physical exertion confounds cardiovascular measures, making them poorly suited for tasks with significant physical loads [14].

2.1.2 Skin Measurements

Dermal measures relate to MW as a response to general Autonomic Nervous System (ANS) stimulation. Electrodermal Activity (EDA), also known as Galvanic Skin Response (GSR), quantifies the electrical resistance of skin. Two electrodes placed an inch apart on the body have a weak current running between them. Changes in measured electrical resistance suggest how the body reacts to external events. The signal is comprised of Skin Conductance Level (SCL) and Skin Conductance Response (SCR), otherwise known as tonic and phasic components. The background tonic response sets a slow-moving baseline that originates from autonomic arousal. Because the tonic SCL differs between individuals, it indicates moderate utility for MW estimation (i.e., tonic components loosely correlate with MW). Phasic SCR comprises faster changing responses to events and thus suggests a stronger MW relationship [9]. Higher frequency and magnitude of phasic peaks correlate to increased MW.

Skin temperature provides some insight to MW, but with temporal and environmental limitations. Slow rise times and delayed event responses make it difficult

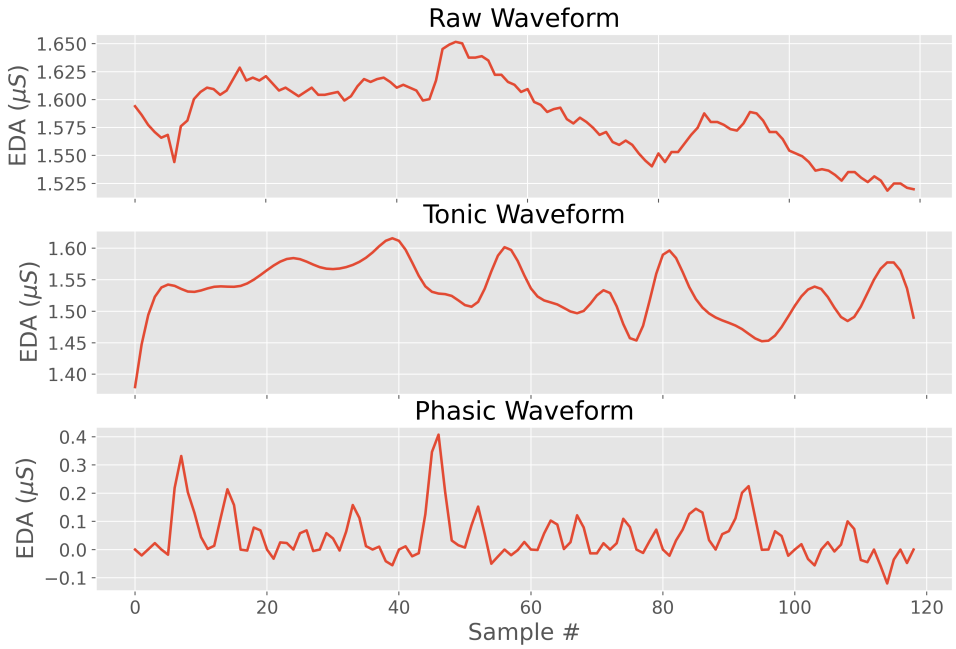


Figure 2.2: Raw EDA measurements decomposed into tonic and phasic waveforms.

to detect changes in MW if MW changes faster than skin temperature reacts. Control of ambient temperature restricts the environment and situations for non-biased skin temperature measurements. Generally, Skin Temperature (ST) decreases in response to MW [21]. However, ST responses to MW may be location dependent [1].

2.1.3 Respiratory Measurements

The Respiration Waveform gives insight into MW through breathing rate and variations [14]. A pressure pad sensor, typically constructed as a chest strap, measures the expansion of the rib cage due to breathing action. In a relaxed state, a human breathes slowly and consistently. However, as MW rises, overall breathing rate increases

along with an increased prevalence of irregular rhythms, quick variations, and cessations [17].

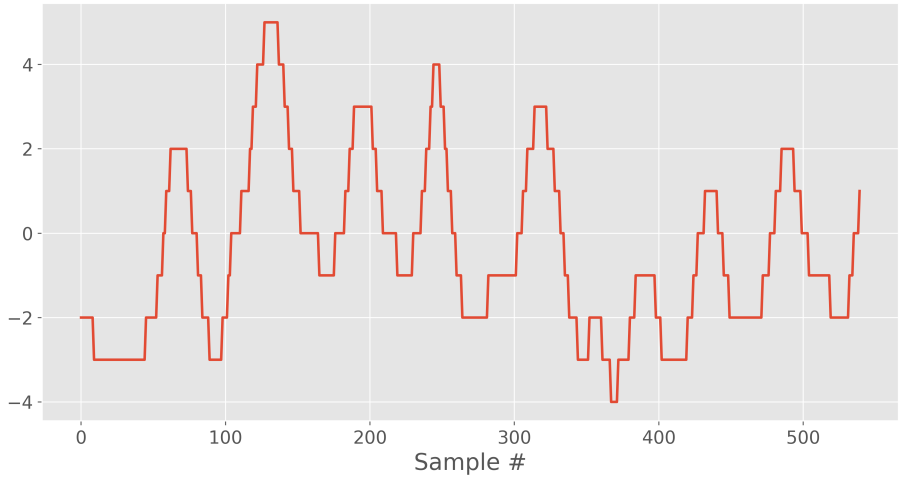


Figure 2.3: Raw Respiration waveform.

2.1.4 Ocular Measurements

Pupillometry measures fluctuations in pupil size and reactivity, which have a direct relationship to MW through the Central Nervous System (CNS) [26]. Several studies have linked blink rate, blink frequency, and pupil diameter to MW [6, 24, 8] Another ocular feature extracts variations in the rate and magnitude of microsaccades, involuntary eye movements that occur during fixation, and have shown to separate different MW levels [19]. One study specifically records pupillometry in VR and found a positive correlation between pupil diameter and subjective task load scores [34]. In virtual environments, VR headsets readily record ocular measurements, but this may not be practical in real-world situations.

Figure 2.4: Plot of raw eye gaze data.

2.1.5 Subjective Measurements

Subjective MW assessments provide a means to validate objective methods of estimating MW. The NASA Task Load Index (NASA-TLX), Bedford [31], and Rating Scale of Mental Effort (RSME) [46] are subjective assessments implemented in MW estimation using physiological measures, with NASA-TLX being the most common [6]. The NASA-TLX consists of six questions where subjects rate their MW on a scale of 1-21. Each question links the given task to a specific category: mental demand, physical demand, temporal demand, performance, effort, and frustration [16]. Results from the NASA-TLX surveys serve as a useful subjective tool to compare and validate MW models. However, subjective MW measurements are susceptible to survey fatigue.

2.2 Gaussian Process Kernels

GP Regression models calculate interpretable estimate functions and confidence intervals based on n-dimensional data with no training data. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution [29]. GPs are defined by a mean function ($m(x)$) and covariance function ($k(x, x')$) (Eq. 2.1) where x and x' are inputs (scalar or vector). In practice, a GP starts with a distribution of possible functions known as priors. The user can decide what type of distribution based on prior knowledge of the data. Given input and output data, the GP calculates a mean function to fit the data as well as its uncertainty at each point along with the function. Understanding the uncertainty of the model provides better prediction interpretations than other ML models. For example, Figure 2.5 shows the mean function and its uncertainty for a single input and single output data, but GPs have the capability to model multivariate data in both the input and output.

$$f(x) \sim \mathcal{GP}\left(m(x), k(x, x')\right) \tag{2.1}$$

Covariance functions, also known as kernels, encode how the GP defines similarity between data. The choice of kernel determines the generalization properties of a GP model dictated by [11]. Each kernel contains hyperparameters such as variance and lengthscale that tune its similarity calculations. The Radial Basis Function (RBF), or Squared Exponential (SE), kernel has become the most common kernel because of its generally sufficient performance to model a wide variety of functions [36, 37]. The RBF

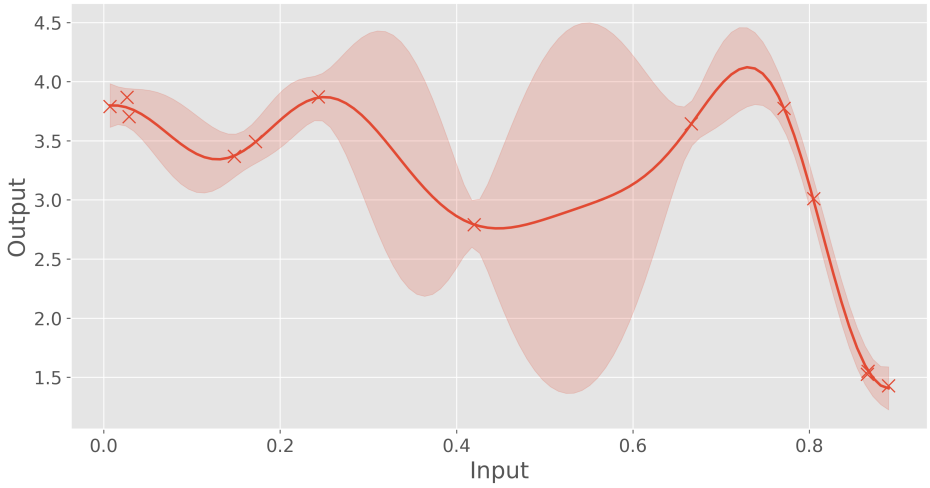


Figure 2.5: Example of a GP mean function estimate showing a $\pm 2\sigma$ confidence interval. Areas of low uncertainty surround measured data points and areas of high uncertainty sit between data.

kernel measures the similarity between two data points by calculating the Euclidean distance (Eq. 2.2). ℓ and σ are the hyperparameters for lengthscale and variance, which determine the reach of the influence and average distance away from the mean respectively. As the distance between data increases, the similarity decreases in the shape of a Gaussian distribution (see Figure 2.6).

$$k_{RBF}(x, x') = \sigma^2 exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) \tag{2.2}$$

For some physical processes, the RBF kernel is considered too smooth, and leads to poor function estimates for data with sharp, rapid changes [38]. While still based on Euclidean distance, the Rational Quadratic (RQ) (Eq. 2.3), Matérn 3/2 (Eq. 2.4), Ornstein-Uhlenbeck (OU) (Eq. 2.5), and Exponential (Eq. 2.6) kernels are

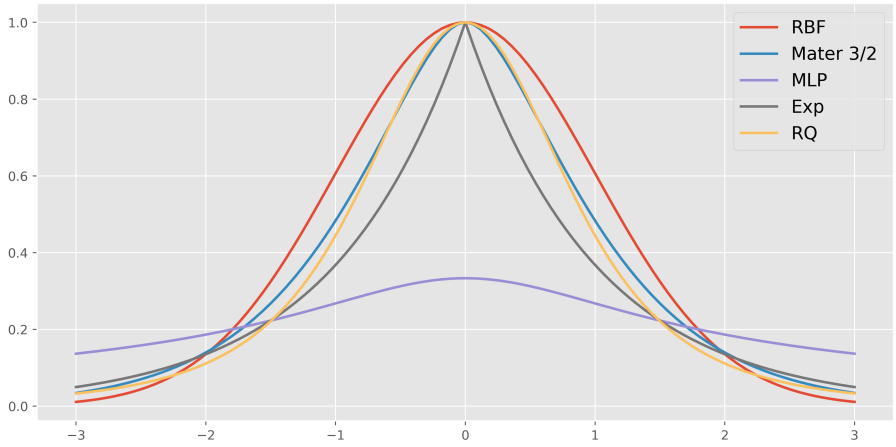


Figure 2.6: Covariance of kernels defined by Euclidean distance. The x-axis represents distance between data points and the y-axis represents similarity. Smaller distances correlate with larger similarity. Generally, kernels with a wider, smooth curve are better at predicting smooth functions.

less smooth and may be better suited compared to the RBF kernel for certain data. Figure 2.6 plots each kernel showing higher data similarity with shorter distances, but as the distance increases, the similarity distribution changes for each kernel. Because of the slight variations of similarity distributions, each kernel produces different resultant function estimates (see Figure 2.7).

$$k_{RQ}(x, x') = \sigma^2 \left(1 + \frac{|x - x'|^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{2.3}$$

$$k_{Mat3/2}(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}|x - x'|}{\ell^2}\right) exp\left(-\frac{\sqrt{3}|x - x'|^2}{2\ell^2}\right) \tag{2.4}$$

$$k_{OU}(x, x') = \sigma^2 exp\left(-\frac{|x - x'|}{2\ell^2}\right) \tag{2.5}$$

$$k_{Exp}(x, x') = \sigma^2 exp\left(-\frac{|x - x'|}{\ell}\right) \tag{2.6}$$

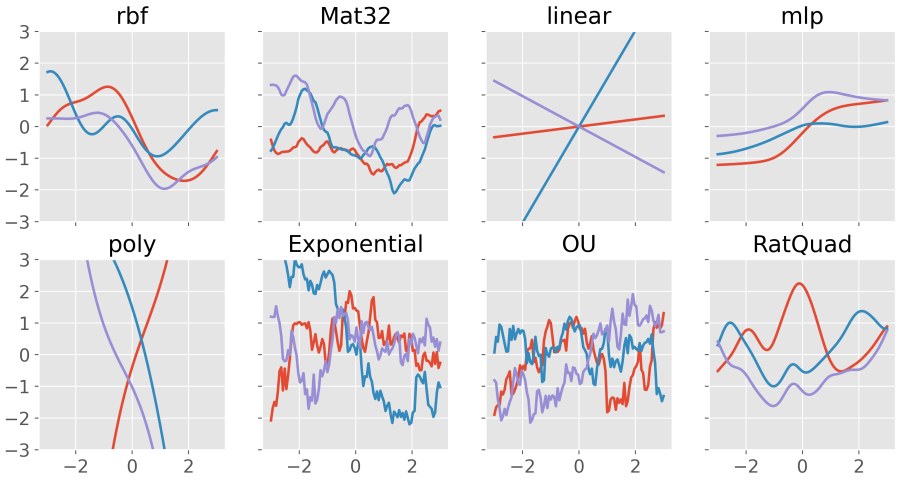


Figure 2.7: Random prior mean functions for several kernels. RBF, MLP, and RQ kernels have smooth predictions while Matérn 3/2, Exponential, and OU produce less smooth functions. The Linear and 3^{rd} degree Polynomial kernels calculate the best fit line or cubic curve respectively.

While many kernels use distance as a metric for similarity, others approach it differently. Linear (Eq. 2.7) and polynomial (Eq. 2.8) kernels serve the same purpose as performing Bayesian linear or polynomial regression. Multilayer Perceptron (MLP) (Eq. 2.9), also known as neural network or arc sine, kernels mimic a feed-forward neural network [28]. Their prior function estimates are shown in Figure 2.7.

$$k_{Lin}(x, x') = \sigma_b^2 + \sigma_v^2(x^\top - c)(x' - c) \tag{2.7}$$

$$k_{Poly}(x, x') = (x^\top x' + \sigma)^d \tag{2.8}$$

$$k_{MLP}(x, x') = \sigma^2 \frac{2}{\pi} asin\left(\frac{\sigma_w x^\top x' + \sigma_b}{\sqrt{\sigma_w x^\top x + \sigma_b + 1}\sqrt{\sigma_w x'^\top x' + \sigma_b + 1}}\right) \tag{2.9}$$

The usefulness of a GP hinges on the model designer's choice of kernel; for multivariate data sets, the choice of kernel is a non-trivial decision. For simple 2-dimensional data, designers can use their intuition to select a kernel based on visualizing the data. However, multivariate data (e.g. multi-modal physiological data) lacks intuitive visual representation to determine the best suited kernel. Previous studies involving physiological data selected RBF for their kernel-based methods [5, 12, 20, 22, 25]. One study used data comprised of features extracted from GSR, ECG, respiration for a Support Vector Machine (SVM) and Extreme Learning Machine (ELM) (kernel-based classifiers) with RBF, linear, and sigmoid kernels [20]. Another study selected the RBF kernel for their GP regression model that predicted cognitive workload based on Electroencephalography (EEG) data [5]. To the best of our knowledge, no study has investigated which kernel best suits each type of data.

2.3 Data Origin

The data set for this thesis comes from Wilson's, et al. study that creates a framework for online multimodal CSE model development and performance prediction [42]. Wilson's framework uses a suite of biosensors to monitor subjects completing tasks in a VR environment while driving a rover on a lunar surface. Subjects sit in a multi-axis motion platform wearing a VR headset, a chest strap sensor, and a wrist sensor. The HTC Vive headset tracks eye movement and pupil size. The Zephyr Bioharness chest strap records ECG and respiration measurements. The Empatica E4 wrist sensor measures EDA and ST. From the raw biosignals, as reported by each sensor, the system performs feature extraction.

Wilson's experiment monitors subject task performance throughout four missions. The objective of each mission is to drive a lunar rover to a designated location while attending to tasks that the rover system requires. The rover system tasks are based on NASA's Multi-Attribute Task Battery II (MATB-II) evaluation, consisting of resource management, system monitoring, tracking, and communications tasks. During the experiment, the VR system records the subject's task performance. The missions vary between high and low difficulties inducing high and low MW. Each subject undergoes two missions at each difficulty in random order with free play time before each mission to allow the body to return to a resting state. Using *Robot Operating System* (ROS), the physiological features and task performance metrics are time synced. NASA-TLX survey responses along with a Random Forest Classifier validated the ex-

periment's ability to modulate between high and low MW.

The selected data set provides a few critical advantages over other data sets. Previous studies derived MW models from a few physiological signals, while Wilson's data contains a more complete suite of signals using passive sensors. Even though other studies created VR environments to immerse subjects, this study incorporates a multi-axis motion platform to deepen the immersive experience and induce more realistic physiological responses. Time synchronicity of biosignal features and performance allows online model development whereas most studies are limited to offline data analysis.

Chapter 3

Problem Statement

The purpose of this chapter is to clarify the problem explored in this study and its scope. Given time-synced physiological data from passive biosensors, we wish to predict human performance for multiple tasks using an array of models with the intention of future online implementations. Because systems involving humans benefit from model prediction interpretability, we model performance with GPs. GPs calculate prediction uncertainty and do not require training data. The input for the GP contains time, principal components of the physiological data, and measured task performance. Time allows the model to put less weight on data that has happened further in the past. To reduce data sparsity and dimensionality, a Principal Component Analysis (PCA) transforms the physiological data into principal components. Although the data contains performance measures for multiple tasks, GPs will only predict one performance metric. The experiment models will output performance predictions, the RMSE of the predictions, and the computation time to predict and update the model at every

time step. This approach assumes tasks can be graded, tasks are done alone, repeated measures are considered independent, and subjects can perform the tasks wearing the required biosensors.

The array of models considered in this study is defined by the different kernels introduced in Section 2.2. Previous research has not performed an in-depth analysis of which kernels yield the best performance predictions for biosignal data. We evaluate each kernel based on their RMSE mean and SEM and mean computation time for each performance metric considered. The results of this study provide a guide for online human performance modeling by detailing appropriate kernel selection and performance calculations.

Chapter 4

Methods

This chapter describes the methods for computing a comparable metric to evaluate kernel selection for Wilson's data set, herein referred to as the VR data set. The first sections provide further details about the VR data set (e.g., the preprocessing procedure) along with an algorithmic overview for simulated online GP modeling. Lastly, we explain the process and rationale for kernel evaluation.

4.1 Data Preprocessing

The VR data set consists of 20 subjects, with each subject completing 4 missions. Before starting the set of four missions, subjects have 5 minutes of free play to establish base physiological states and get a feel for operating the rover system. After the initial free play, the first mission starts. Upon completion of the mission, the subjects complete a NASA-TLX survey. Again, the subject has a free play session, but for only 3 minutes. In total the subject has four missions with a free play session before

$$\left[\begin{array}{c|ccc|ccc} t_1 & f_{1,1} & \cdots & f_{83,1} & p_{1,1} & \cdots & p_{4,1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ t_n & f_{1,n} & \cdots & f_{83,n} & p_{1,n} & \cdots & p_{4,n} \end{array}\right]$$

Figure 4.1: Shape of data from one mission. At each time step t there are 83 features f and 4 performance metrics p.

each one (the first being 5 minutes and the rest 3 minutes). The GP models examine each of the 4 missions independently with its implications described in Section 5.2.2. Data from each mission consists of 83 physiological features and 4 performance metrics at each time step (see Figure 4.1)

Figure 4.2: Performance measures.

Performance measures track resource management, system monitoring, track-

ing, and communications based on the MATB-II. While the data contains metrics for multiple tasks within each category, this study considers one metric per category for GP kernel evaluation (see Figure 4.2). The resource management metric describes the likelihood of reaching the objective point by computing the ratio of remaining oxygen to remaining distance to the goal. Engine temperature monitoring represents system monitoring. The subjects' ability to maintain a strong communication signal with the onboard antenna falls under the tracking task, which requires the user to monitor their heading with respect to a fixed point in the environment. Response time to communication channel requests tracks communication task performance. While a subject's performance from mission to mission can improve from learning, the randomized order of difficulty helps remove bias. The purpose of this study revolves around determining the appropriate kernel for physiological feature data, not the method of measuring performance.

GP models are prone to high uncertainty predictions where data is sparse (see Figure 2.5). To reduce data sparsity, the simulation free play data undergoes PCA. The PCA helps reduce sparsity and serves as a dimensionality reduction by calculating the principal components of the data that have an associated variance explained. Closely grouped data, along with fewer dimensions, reduce the computation time for GP model optimization. Implementing the PCA class from the *sklearn* Python library, we set the percentage of desired variability explained to 95%, which returns 16 principal components on average. The rationale for selecting 95% variability explained comes from finding the average number of principal components required to reach each percentage of

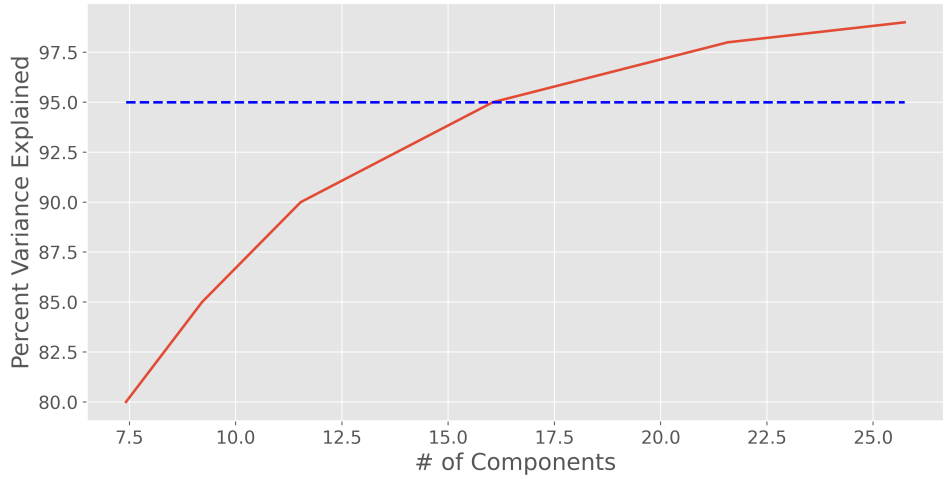


Figure 4.3: Plot of percentage of variance explained by the average number of principal components with 95% as the selected percentage.

variability explained. Figure 4.3 demonstrates that increasing the variability explained from 95% to 98% requires 31.2% more components, and the small amount of added variability explained does not justify the increase in computation time from additional principal components. The PCA reduces the 83 input dimensions from the original data set by 80%. With fewer input data, computation time for calculating GP models diminishes, which better suits online applications.

Including a PCA step also provides insight into which sources of data have the largest variability and potentially more useful information. While greater variation does not exactly equate to more useful information, it does provide insight into physiological data's connection to mental workload. Chapter 5 discusses PCA interpretation in further detail.

4.2 Gaussian Process Model Development

In order to simulate online GP modeling, the model predicts performance and updates at each time step. Algorithm 1 provides a high level model overview, where x, $\boldsymbol{x}$, X, and $\boldsymbol{X}$ represent scalars, vectors, program objects, and matrices respectively. The algorithm runs once per mission and starts by creating a principal component transform function T_{form} based on free play data $\boldsymbol{A}$. The input data $\boldsymbol{B}$ appends the physiological features at each time step, $\boldsymbol{b_i}$, until the input reaches the desired size, determined by t_0, to initialize the GP model. For this experiment $t_0 = 10$, meaning that the initial GP model starts with 10 seconds of data. Physiological data undergoes a transformation into principal components $\boldsymbol{B'}$. For each kernel K_i and performance metric $\boldsymbol{P_i}$ the algorithm creates or updates a GP with the transformed data. The kernels enable Automatic Relevance Determination (ARD) and use the number of principal components as active dimensions. ARD implicitly determines the relevance of each input by creating a length-scale parameter for each input.

Algorithm 1 Top Level Algorithm

```
Set parameters: t_0, K
T_form ← getDimRedux(A)
while running do
    B ← B.append(b_i)
    while t_i ≥ t_0 do
        B' ← T_form(B)
        for K do
            for P do
                μ̂_{t+1}, σ̂_{t+1}, ε_t ← gpAnalysis(B', P_i, K_i)
            end for
        end for
    end while
end while
```

Algorithm 2 functions as a human performance predictor, with time, principal components, measured performance, and kernel as the inputs. Once initialized, the model computes the predicted performance mean value $\hat{\mu}_{t+1}$ and its variance $\hat{\sigma}_{t+1}$ for the next time point based on the current mean function. The model evaluates the prediction by calculating prediction RMSE. Lastly, Line 7 updates H, the hyperparameters, of the chosen K to be used at the next time step. To ensure model trustworthiness, we validated Algorithm 2 by analyzing model evolution with increased data (see Appendix A). If implemented online, the algorithm would receive new data at 1Hz (the frequency of data collection from the VR study), but with the entire data set present, this study neglects time between incoming data and instead presents data sequentially to mimic an online environment.

Algorithm 2 Gaussian Process Algorithm

$gpAnalysis(\boldsymbol{B}, \boldsymbol{P}$,K)

1: **if** $not(initialized)$ **then**
2: $\quad GP \leftarrow GP.initialize(\boldsymbol{B}, \boldsymbol{P}, K)$
3: **else**
4: $\quad \hat{\mu}_{t+1}, \hat{\sigma}_{t+1} \leftarrow GP.predict(\boldsymbol{B}, \boldsymbol{P})$
5: $\quad \epsilon_t \leftarrow modelEval(p_t, \hat{p}_1)$
6: $\quad H \leftarrow GP.update(\boldsymbol{B}, \boldsymbol{P})$
7: **end if**
8: $return\ \hat{p}_{t+1}, \hat{v}_{t+1}, \epsilon_t$

Since kernel performance may vary between performance metrics, each metric has its own GP. The algorithms can estimate functions with multivariate data, but each GP only takes one performance measure into account (discussed further in Section 5.2.2). Thus, $\boldsymbol{B}$ is multivariate, but $\boldsymbol{P}$ is one dimensional. Computing performance estimate functions for each performance measure allows comparisons between kernel

performance that a multivariate performance output may hide.

4.3 Kernel Evaluation

The choice of kernel determines the generalization properties of a GP model [11]. Since the purpose of this study's GP models are to predict performance, we evaluate kernel selection by model prediction accuracy. This experiment calculates model prediction RMSE at each time step. By normalizing the RMSE with the average true performance value, the errors serve as a comparable model performance metric across each kernel and task performance metric.

The kernels introduced in Chapter 2 are the candidates used in this study: RBF, RQ, Matérn 3/2, OU, Exponential, Linear, 3^{rd} degree Polynomial, and MLP kernels (see Figure 2.7). This study aims to find a kernel that outperforms the most commonly implement kernel, RBF. Matérn 3/2, OU, and Exponential kernels produce less smooth estimates compared to RBF, while RQ and MLP compute smoother estimates. Linear or cubic fits may produce a lower error and would add the benefit of drastically decreasing computation time. For each kernel, the model undergoes Algorithm 2 and the resultant average RMSEs are compared. Overall, the experiment analyzes 8 kernels over 20 subjects' data which includes 4 performance measures for all 4 missions for a total of 2,560 models.

Chapter 5

Results & Discussion

The following sections report the findings and interpretations of the GP human performance prediction. Section 5.1 summarizes the results of the study. Section 5.2.1 describes a heuristic for understanding which biosignals supply greater information for human performance modeling. Section 5.2.2 provides a visualization for GP predictions, draws conclusions from kernel performance, and suggests further work.

5.1 Results

5.1.1 Principal Component Analysis

PCA calculates the principal components of a data set, along with the contribution of each data type to the component's variability, referred to as loadings. Each principal component contains a percentage of the original data information. Figure 5.1 represents each physiological feature's contribution to the total variability of each

component. The figure orders principal components along the x-axis from highest to lowest percent variability explained. The first principal component has ECG, EDA, and pupillometry frequency domain features comprising most of the variation. Pupillometry time-domain features hold the majority of variability in the second principal component.

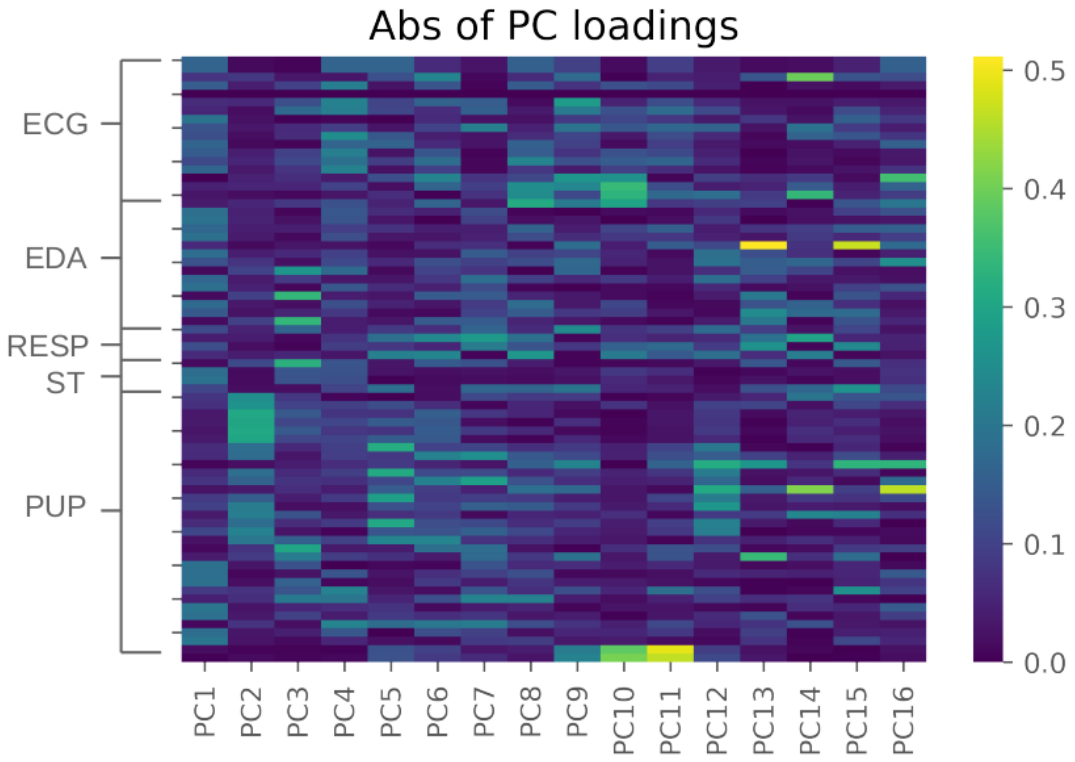


Figure 5.1: Heat map of principal components and the amount of variation explained per physiological feature. Features are grouped by source sensor on the left

5.1.2 Gaussian Process Performance Prediction Surfaces

The methods described in Chapter 4, created a GP for each mission, performance measure, and kernel totaling 2,560 models and calculated RMSEs for performance predictions made at each time step. Before reviewing the kernel performance, we can

visualize a simplified version of the problem to gain better intuition of the results. Figure 5.2 shows one principal component and performance metric over time. By reducing the multivariate models to a single principal component and performance metric (i.e. the data input and output) over time and predicting along the principal component axis, the GP predictions create a surface. The plots overlay the measured values on the prediction performance surfaces. While the mean and SEM of the RMSEs (discussed later in Section 5.2.2) measure the accuracy of the models, the surfaces provide further insight. Figure 5.2 shows the surfaces from the RBF, Linear, Matérn 3/2, and MLP kernels. The RBF, Matérn 3/2, and MLP kernels produce similar surfaces, with MLP having the smoothest predictions. The Linear kernel appears to have accurate local estimations but drastically deviates from the other models away from measured values.

5.1.3 Kernel Root Mean Square Error

Given the RMSEs from the experiment, this section shows the kernel accuracy across the four performance metrics. For each mission's performance metric, we calculate the prediction RMSE mean and SEM. RMSE was selected over Mean Absolute Error (MAE), because RMSE is more sensitive to larger errors. In this study, the model predictions are 1D, so the two are equivalent. However, if the model were to predict n-dimensions, it may matter a great deal. To allow comparisons between performance metrics, the means and SEMs are first normalized by the mean measured performance (Equations 5.1 and 5.2). Since each mission's length of time varies, we calculate the weighted mean RMSE (Equation 5.3) and pooled SEM (Equation 5.4) across missions

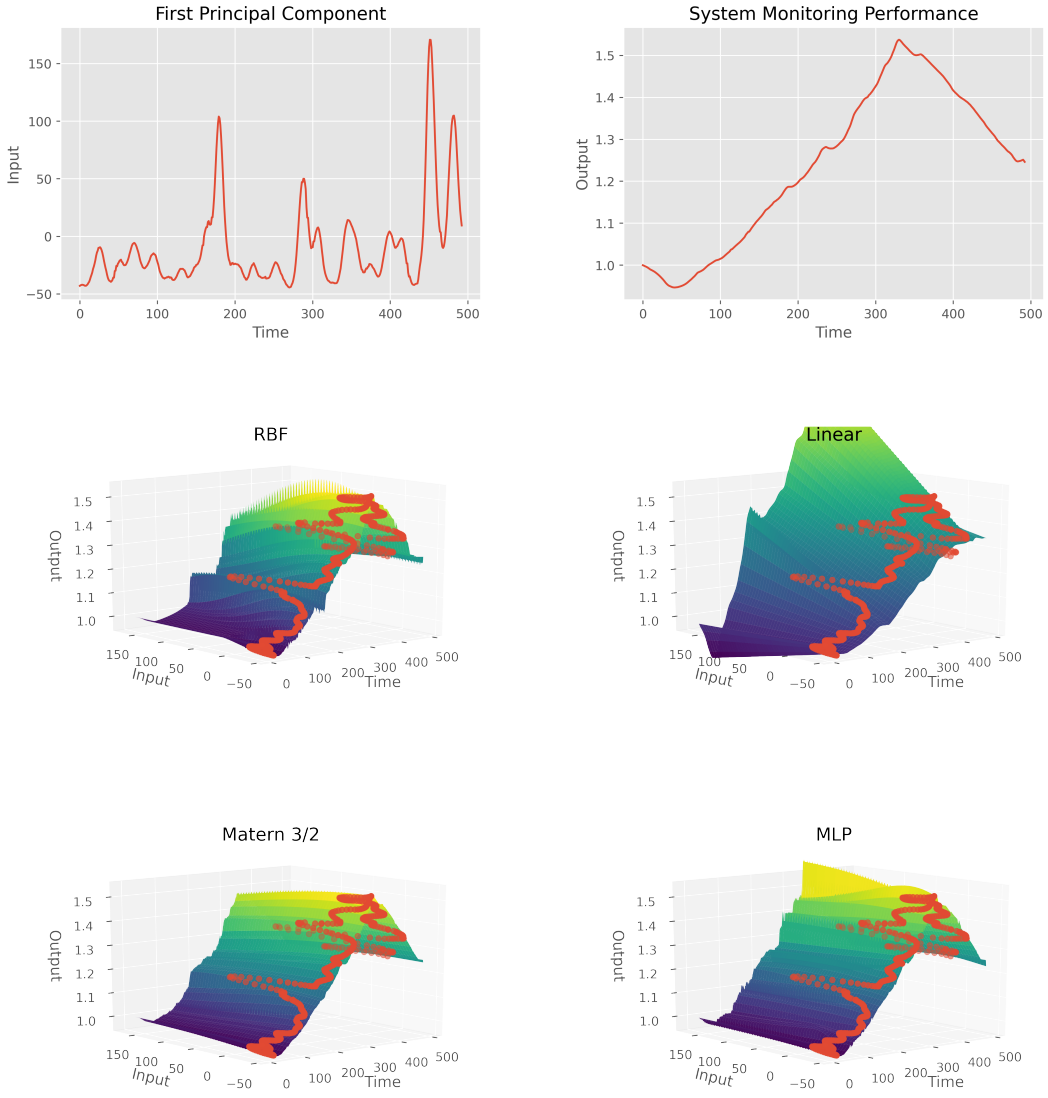


Figure 5.2: Surface Performance Predictions.

for each kernel and performance metric. Table 5.1 shows the resulting RMSE means and SEM. For all performance metrics, the RQ and MLP kernels yielded the lowest RMSE and SEM. Table 5.2 describes the RMSEs normalized by the lowest RMSE for each performance metric. It is important to also recognize that not all models improved

accuracy with more time and data (see Figure 5.4).

$$RMSE_{normalized} = \frac{\sqrt{\frac{1}{n}\sum_{i}^{n}(y_p - y_i)^2}}{\bar{y}} \tag{5.1}$$

$$SEM_{normalized} = \frac{SEM}{\bar{y}} \tag{5.2}$$

$$RMSE_{mean} = \frac{n_1 RMSE_1 + n_2 RMSE_2 + ... + n_k RMSE_k}{n_1 + n_2 + ... + n_k} \tag{5.3}$$

$$SEM_{pooled} = \sqrt{\frac{(n_1 - 1)SEM_1^2 + (n_2 - 1)SEM_2^2 + ... + (n_k - 1)SEM_k^2}{n_1 + n_2 + ... + n_k - k}} \tag{5.4}$$

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	0.36 ± 0.020	0.19 ± 0.031	0.28 + 0.027	0.09 ± 0.004
RQ	**0.25 ± 0.018**	**0.06 ± 0.017**	**0.14 ± 0.018**	**0.03 ± 0.002**
Mat 3/2	0.36 ± 0.020	0.20 ± 0.031	0.28 + 0.027	0.09 ± 0.004
Exp	0.36 ± 0.020	0.22 ± 0.032	0.30 ± 0.027	0.10 ± 0.004
OU	0.36 ± 0.020	0.22 ± 0.032	0.30 ± 0.027	0.10 ± 0.004
Linear	0.39 ± 0.023	0.14 ± 0.025	0.31 ± 0.025	0.06 ± 0.002
Cubic	0.40 ± 0.038	0.08 ± 0.025	0.36 ± 0.218	0.06 ± 0.024
MLP	**0.25 ± 0.019**	**0.05 ± 0.020**	**0.15 ± 0.020**	**0.03 ± 0.002**

Table 5.1: RMSE means and SEM across each kernel and performance measure. Bolded values represent the lowest mean and SEM. RQ and MLP yielded the best performance across all performance metrics.

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	1.44	3.62	2.02	3.06
RQ	1.01	1.15	1.00	1.02
Mat 3/2	1.44	3.72	2.04	3.06
Exp	1.46	4.07	2.17	3.29
OU	1.46	4.07	2.17	3.29
Linear	1.57	2.66	2.22	1.96
Cubic	1.64	1.57	2.60	2.20
MLP	1.00	1.00	1.12	1.00

Table 5.2: RMSE means as a ratio of RMSE over the lowest RMSE for each performance metric.

5.1.4 Computation Time

With the motivation to implement performance predictions online, we recorded the computation time to predict performance, update the model, and optimize hyperparamters at each time step (i.e. time to complete Algorithm 2). As the model incorporates more data, the computation time and its variation increases linearly (see Figure 5.3). Table 5.3 shows the mean and SEM computation times for each kernel and performance metric. Calculations for means and SEM used Equations 5.3 and 5.4 due to differing lengths of missions.

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	4.90 ± 0.20	5.95 ± 0.26	5.36 ± 0.25	3.58 ± 0.14
RQ	5.66 ± 0.22	8.24 ± 0.42	4.54 ± 0.19	2.26 ± 0.08
Mat 3/2	4.94 ± 0.20	5.82 ± 0.25	4.59 ± 0.18	3.58 ± 0.13
Exp	4.83 ± 0.18	4.87 ± 0.19	4.23 ± 0.15	3.74 ± 0.13
OU	4.83 ± 0.18	4.87 ± 0.19	4.23 ± 0.15	3.74 ± 0.13
Linear	3.11 ± 0.15	4.52 ± 0.18	3.55 ± 0.15	3.38 ± 0.13
Cubic	1.52 ± 0.09	4.91 ± 0.27	3.15 ± 0.15	3.73 ± 0.16
MLP	4.23 ± 0.21	6.35 ± 0.34	4.62 ± 0.20	2.93 ± 0.13

Table 5.3: Computation time means and SEM.

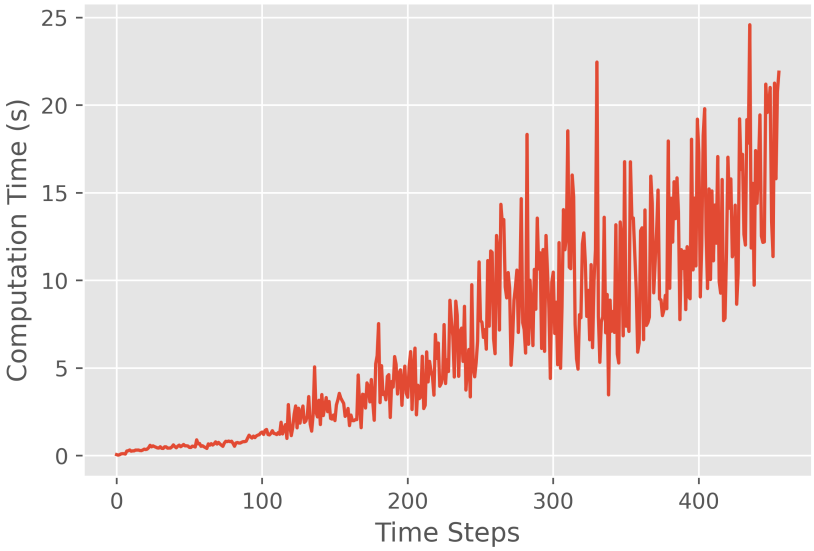


Figure 5.3: Plot of computation time with increased data.

Because computation time largely depends on hardware, Table 5.4 reports values as ratios of mean computation time over the shortest computation time. The cubic kernel computed the model updates fastest for tracking, thus it has a value of 1.00. Whereas, RBF for tracking took 3.22 times the computation time. MLP, one of the higher accuracy kernels, takes less time to compute for all but resource management compared to RBF. RQ, the other high-performing kernel, only computes faster for communications and system monitoring. While MLP and RQ outperform the other kernels in accuracy for all metrics, they do not outperform in computation time.

Tables 5.5 and 5.6 describes computation time as comparisons within each performance metric and kernel respectively. For tracking, resource management, and communications linear and cubic kernels compute the fastest. Between kernels, they

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	3.22	3.91	3.52	2.35
RQ	3.72	5.41	2.99	1.49
Mat 3/2	3.25	3.83	3.02	2.39
Exp	3.17	3.20	2.78	1.46
OU	3.17	3.19	2.77	2.46
Linear	2.05	2.97	2.33	2.22
Cubic	**1.00**	3.23	2.07	2.45
MLP	2.78	4.18	3.04	1.92

Table 5.4: Computation times as a ratio of mean time over lowest overall mean time.

computed system monitoring quickest except for the linear and cubic kernels which favored tracking.

5.2 Discussion

5.2.1 Principal Component Analysis Interpretation

An added bonus of PCA is additional insight into which data features contain the most variability. Figure 5.1 shows a heatmap of the principal components and the contribution of each physiological feature to the principal component's variability explained. By taking the first two or three components, we can draw coarse conclusions about which features contain the most information. In Figure 5.1, ECG, EDA, and pupillometry features make up the majority of the variation in the first component. Pupillometry contributes the most in the first three components, which suggests that ocular measurements provide more information than other biosignals.

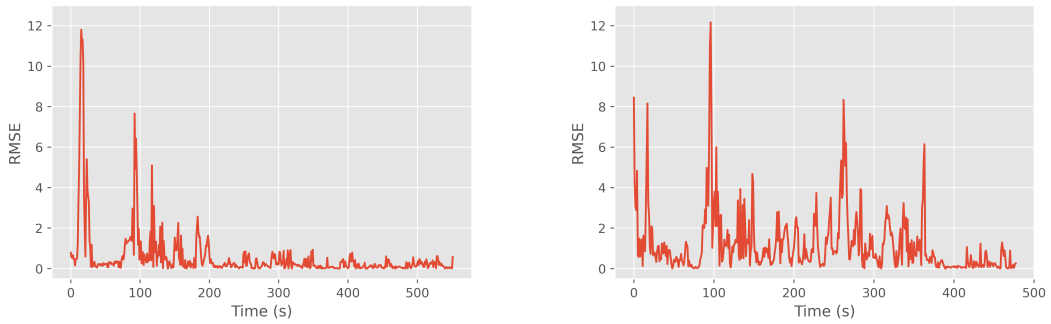
While larger variability in signal does not exactly mean more useful infor-

mation, PCA yields a heuristic into which signals may prove more informative. The number of features from each physiological signal may cause redundancy. For example, the second component in Figure 5.1 shows a lot of variation in ocular features. Some of those features may give the same information. Narrowing the features to ones that supply the most useful information could decrease the number of principal components, lead to faster and more accurate estimates, and reduce the number of required sensors to acquire data.

5.2.2 Kernel Performance

While the weighted mean and SEM of RMSE represent the accuracy and precision of the predictions, they do not give the evolution of the prediction error over time. Figure 5.4 gives two examples of the RMSE over time. Figure 5.4a represents expected decay of RSME over time because with more data, the model predictions should improve. Figure 5.4b shows the RMSE for the same performance metric and kernel, but different subject. Instead of the error decreasing over time, the mean error and spread do not decay. One potential cause of no error decay could stem from subject learning. The purpose of including time in the model is to allow for the model to neglect older measurements, assisting in modeling learning to a degree. Applying a limit to the amount of data in the model, something not done here, could better force the model to forget older data that may not be relevant because of subject learning. Another cause, related to learning, may derive from treating each mission independently. As subjects complete each mission, they learn and adapt. During the learning process, the

subject's physiological responses may differ from post-learning in response to the same task demands. A model designed to account for repeated measures may improve model performance.



(a) (b)

Figure 5.4: Plots RMSE over time for two subjects for the same kernel and performance metric. Plot (a) shows a decay of RSME over time while plot (b) does not decay.

Table 5.3 presents RQ and MLP as the most accurate kernels for modeling performance from physiological data. RQ and MLP may outperform other kernels because they have smoother priors compared to RBF. The less smooth kernels, Matérn 3/2, Exponential, and OU, yielded approximately the same prediction accuracy and spread as RBF. The results suggest that a smoother kernel may better suit human physiological and performance data sets.

The simple Linear and Cubic kernels outperformed RBF for some performance metrics. While a Linear fit model may perform better than other kernels for resource management and system monitoring, Figure 5.2 shows that a Linear model may only have good predictions locally. Because the experiment only predicts one second in the future, simple or smooth models may suffice. Future work should evaluate kernel

selection based on predicting multiple observations in the future.

The type of estimated performance metric contributes to the GP model prediction accuracy. The rankings of each kernel based on lowest mean (Table 5.1), generally stays the same across all performance metrics. However, the range of mean and SEM changes between performance metrics with the highest errors coming from tracking and lowest from system monitoring. Because each performance metric yields a different accuracy, model designers may need to tune their model(s) to each metric or even change how to calculate the metrics.

By creating GP models for each performance metric, we can understand the relationship between physiological responses and each metric. Many findings from this section would not have been apparent had the experiment models used a multivariate output. Individual models allowed differences in kernel performance for each metric to emerge. The results show that the RQ and MLP kernels perform best across each metric. Knowing that performance metrics do not determine which kernel performs best, the future GP models could estimate all performance metrics as a multivariate output using one kernel.

Differing accuracy and spread across performance metrics may result from different physiological responses to each demand. The higher mean error for tracking performance predictions may stem from the difficulty in tracking quick changes based on slower physiological responses. As shown in Figure 4.2, antenna accuracy, or tracking, rapidly changes. Interestingly, for rapidly changing performance metrics, less smooth kernels did not outperform smooth kernels. While some physiological responses slowly

change (e.g. heart rate, breathing rate, skin temperature), faster responses (e.g. pupil size, eye gaze) may better predict tracking. Perhaps a multimodal approach for all performance metrics does not yield the best models.

Resource management mean RMSE had the highest spread (Table 5.1), most likely due to poor performance metric calculations. The resource management metric, or O^2 performance (see Figure 4.2), exponentially increases with good performance. As the subject approaches the mission objective, the denominator of the metric, distance remaining, goes to zero. This causes an unusually high error towards the end of a successful mission, which may explain the high SEM. However, even with high spreads, the smooth kernels' accuracy exceeded all but system monitoring predictions. Unlike tracking, resource management changes slowly over time, which suits RQ and MLP kernels. To lower mean and standard errors, the resource management metric needs bounds that prevent exponential changes.

Communication performance predictions also yielded high SEM (Table 5.1). The communications metric calculates the subject time response to requests, but that time changes step-wise and its changed value depends on the occurrence of requests (see Figure 4.2). The subject most likely responds between the time of request and response. The metric holds that response time until the next request, which may not correlate well with MW. In order to lower mean error and spread, the model should estimate performance during the window of time that the subject responds to the request.

The system monitoring metric reflects the rover engine temperature and had the lowest mean and spread for all kernels (Table 5.1). The level of accuracy and

precision suggests that physiological responses to the demand of system monitoring correlate well. As a background task, it may not cause much response, but during critical portions of the mission (e.g. scaling a hill) the engine temperature increase probably induces a large response. The magnitude of response between each task demand may also contribute to the lack of accuracy in other metrics.

For online modeling, computation time contributes to overall kernel performance. Table 5.5 compares computation time by performance metric. It is no surprise that for three of the four metrics, Linear and Cubic function fits compute faster. However, for the lowest prediction error metric, system monitoring, RQ computes fastest. Ignoring the Linear and Cubic kernels, MLP has the lowest computation time for tracking, and OU and Exponential have the lowest for resource management and communications. RQ and MLP time suggests they are not only good candidates based on accuracy but also computation time as a single choice kernel to estimate all performance metrics. To meet the speed required for online modeling, the amount of data considered in a model needs a reduction. By setting a limit, or running window, of data size (i.e. model memory), the computation time will decrease, but further study needs to verify the effects on model accuracy.

The most accurate models also compute the fastest. Comparing computations by kernel choice, all but Linear and Cubic compute fastest for system monitoring (Table 5.6). If estimating functions for well correlated data requires less computation time (as Tables 5.1 and 5.6 suggest), tuning the right physiological features to each task demand will better correlate data and decrease computation time. It is important to

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	3.22	1.32	1.70	1.58
RQ	3.72	1.82	1.44	**1.00**
Mat 3/2	3.25	1.29	1.46	1.58
Exp	3.17	1.08	1.34	1.65
OU	3.17	1.07	1.34	1.65
Linear	2.05	**1.00**	1.13	1.49
Cubic	**1.00**	1.09	**1.00**	1.65
MLP	2.78	1.41	1.47	1.29

Table 5.5: Computation times as a ratio of mean time over the lowest mean time for each performance measure (column).

note that the largest contribution to computation time comes from optimizing the kernel hyperparameters. As described in Algorithm 2, model optimization occurs at every time step. Reducing the frequency of model optimization would decrease computation time, but the amount of reduction to not decrease model performance needs further study.

	Performance Metric			
Kernel	Tracking	Resource Mgmt	Comms	Sys Monitoring
RBF	1.37	1.66	1.50	**1.00**
RQ	2.50	3.64	2.01	**1.00**
Mat 3/2	1.38	1.63	1.28	**1.00**
Exp	1.29	1.30	1.13	**1.00**
OU	1.29	1.30	1.13	**1.00**
Linear	**1.00**	1.45	1.14	1.09
Cubic	**1.00**	3.23	2.07	2.45
MLP	1.44	2.17	1.58	**1.00**

Table 5.6: Computation times as a ratio of mean time over the lowest mean time for each kernel (row).

5.3 Future Work

Further investigations should aim to improve predictions by adjusting the number and type of physiological features, improving performance metrics, and conducting a deeper analysis of kernel selection. Given the physiological insight from PCA, a study to verify what biosignal features contain the most useful information, could improve performance predictions. In a practical sense, a study dedicated to only using signals from some of the sensors would help reduce costs, complexity, and sensors required. The possibility of tuning models for each performance metric by only considering features that correlate well with that metric could prove useful. For example, estimating rapidly changing tracking performance may require physiological features that can change and react quickly (e.g. eye movement). Ultimately, the question of which features offer the most information and their correlation to different task performance needs more exploration.

Based on the findings in Section 5.2.2, we better understand what makes a good performance metric. The communications metric would improve by limiting model data and predictions to the time between requests and response so that the subjects' physiological signals reflect that demand. Also, bounding performance metrics will help shrink the spread of error, especially for resource management.

While the experiment resulted in two kernels with higher accuracy, future work needs to verify this selection for predictions made further in the future, with limited data, and for multivariate outputs. By making predictions greater than one

second ahead, the kernels that previously performed equally may separate. In addition to changing prediction methods, researchers can explore combinations of kernels by adding or multiplying them together. Combining kernels increases kernel complexity and may increase computation time because of the added hyperparameters to optimize, but improved predictions could justify the additional time cost. In order to improve computation time, the amount of data held within the GP needs to have a limit, or the frequency of model optimization needs to decrease. Creating a window of data will help keep computation time under a desired threshold, but the effects on accuracy deserve further study. In the long term, model designers should consider a combined model where a single GP predicts all of the performance metrics. A holistic model may better reflect the complexity of MW in performance.

Chapter 6

Conclusion

The results of this study show that smooth kernels, compared to RBF, better suit human performance models based on physiological data. RQ and MLP kernels yield lower mean RMSE, SEM and computation time compared across all performance metrics considered. Our PCA interpretation and analysis of mean RMSE between performance metrics provide insight into understanding physiological data and its correlation to MW and human performance, which is key to improving confidence in MW models. By improving MW models, we decrease the likelihood of models leading to poor safety, performance, and task completion rate in high-risk, high operational cost scenarios. The comparison of computation time for model prediction and optimization points to further refinement to allow online modeling. The combination of improved MW and performance modeling and necessary adjustments for online modeling help form the basis for better HR teaming and future adaptive, closed-looped control systems.

Bibliography

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017.

[2] Polina Andrievskaia, Kathleen Van Benthem, and Chris M Herdman. Neural correlates of mental workload in virtual flight simulation. In *International Conference on Human-Computer Interaction*, pages 521–528. Springer, 2020.

[3] Pietro Aricò, Gianluca Borghini, Ilenia Graziani, Fumihico Taya, Yu Sun, Anastasios Bezerianos, Nitish V Thakor, Febo Cincotti, and Fabio Babiloni. Towards a multimodal bioelectrical framework for the online mental workload evaluation. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3001–3004. IEEE, 2014.

[4] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013.

[5] Matthew S Caywood, Daniel M Roberts, Jeffrey B Colombe, Hal S Greenwald, and Monica Z Weiland. Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks. *Frontiers in human neuroscience*, 10:647, 2017.

[6] Rebecca L Charles and Jim Nixon. Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, 74:221–232, 2019.

[7] Lan-lan Chen, Yu Zhao, Peng-fei Ye, Jian Zhang, and Jun-zhong Zou. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, 85:279–291, 2017.

[8] Melissa Patricia Coral. *Analyzing cognitive workload through eye-related measurements: A meta-analysis*. PhD thesis, Wright State University, 2016.

[9] Dick De Waard and KA Brookhuis. The measurement of drivers' mental workload. 1996.

[10] Yi Ding, Yaqin Cao, Vincent G Duffy, Yi Wang, and Xuefeng Zhang. Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, 63(7):896–908, 2020.

[11] David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[12] Hernán F. García, Mauricio A. Álvarez, and Álvaro A. Orozco. Gaussian process dynamical models for multimodal affect recognition. In *2016 38th Annual In-*

ternational Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 850–853, 2016.

[13] Sukeshini Grandhi and Quentin Jones. Technology-mediated interruption management. *International Journal of Human-Computer Studies*, 68(5):288–306, 2010.

[14] Mariel Grassmann, Elke Vlemincx, Andreas Von Leupoldt, Justin M Mittelstädt, and Omer Van den Bergh. Respiratory changes in response to cognitive load: a systematic review. *Neural plasticity*, 2016.

[15] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 301–310, 2010.

[16] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[17] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[18] Jens Kohlmorgen, Guido Dornhege, Mikio Braun, Benjamin Blankertz, Klaus-Robert Müller, Gabriel Curio, Konrad Hagemann, Andreas Bruns, Michael Schrauf,

Wilhelm Kincses, et al. Improving human performance in a real operating environment through real-time mental workload detection. *Toward brain-computer interfacing*, 409422:409–422, 2007.

[19] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, 13(9):e0203629, 2018.

[20] Lan lan Chen, Yu Zhao, Peng fei Ye, Jian Zhang, and Jun zhong Zou. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, 85:279–291, 2017.

[21] Charlotte Larmuseau, Jan Cornelis, Luigi Lancieri, Piet Desmet, and Fien Depaepe. Multimodal learning analytics to investigate cognitive load during online problem solving. *British Journal of Educational Technology*, 51(5):1548–1562, 2020.

[22] Jie Liang, Zhengyi Shi, Feifei Zhu, Wenxin Chen, Xin Chen, and Yurong Li. Gaussian process autoregression for joint angle prediction based on semg signals. *Frontiers in Public Health*, 9, 2021.

[23] Tiffany Luong, Nicolas Martin, Anais Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lécuyer. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 425–437. IEEE, 2020.

[24] Kevin Mandrick, Vsevolod Peysakhovich, Florence Rémy, Evelyne Lepron, and Mickaël Causse. Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biological psychology*, 121:62–73, 2016.

[25] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical review letters*, 100(14):144103, 2008.

[26] Gerhard Marquart, Christopher Cabrall, and Joost de Winter. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3:2854–2861, 2015.

[27] Niloofar Momeni, Fabio Dell'Agnola, Adriana Arza, and David Atienza. Real-time cognitive workload monitoring based on machine learning using physiological signals in rescue missions. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3779–3785, 2019.

[28] Guofei Pang, Liu Yang, and George Em Karniadakis. Neural-net-induced gaussian process regression for function approximation and pde solution. *Journal of Computational Physics*, 384:270–288, 2019.

[29] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[30] Annika Rieger, Regina Stoll, Steffi Kreuzfeld, Kristin Behrens, and Matthias Weip-

pert. Heart rate and heart rate variability as indirect markers of surgeons' intraoperative stress. *International archives of occupational and environmental health*, 87(2):165–174, 2014.

[31] Alan H. Roscoe and George A. Ellis. A subjective rating scale for assessing pilot workload in flight: A decade of practical use. 1990.

[32] Kilseop Ryu and Rohae Myung. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35(11):991–1009, 2005.

[33] Sebastian Sauer, Harald Walach, Stefan Schmidt, Thilo Hinterberger, Siobhan Lynch, Arndt Büssing, and Niko Kohls. Assessment of mindfulness: Review on state of the art. *Mindfulness*, 4(1):3–17, 2013.

[34] Maximilian Schwalm, Andreas Keinath, and Hubert D Zimmer. Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation*, 1986:1–13, 2008.

[35] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5, 2017.

[36] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[37] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[38] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[39] Pieter Vanneste, Annelies Raes, Jessica Morton, Klaas Bombeke, Bram B Van Acker, Charlotte Larmuseau, Fien Depaepe, and Wim Van den Noortgate. Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, pages 1–19, 2020.

[40] Chixiang Wang and Junqi Guo. A data-driven framework for learners' cognitive load detection using ECG-PPG physiological feature fusion and XGBoost classification. *Procedia computer science*, 147:338–348, 2019.

[41] Glenn F Wilson and Christopher A Russell. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human factors*, 49(6):1005–1018, 2007.

[42] Robert L Wilson, Daniel Browne, Jonathan Wagstaff, and Steve McGuire. A virtual reality simulation pipeline for online mental workload modeling. *arXiv preprint arXiv:2111.03977*, 2021.

[43] Mark S Young, Karel A Brookhuis, Christopher D Wickens, and Peter A Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.

[44] Xiao Zhang, Yongqiang Lyu, Xin Hu, Ziyue Hu, Yuanchun Shi, and Hao Yin. Evaluating photoplethysmogram as a real-time cognitive load assessment during game playing. *International Journal of Human–Computer Interaction*, 34(8):695–706, 2018.

[45] Bin Zheng, Xianta Jiang, Geoffrey Tien, Adam Meneghetti, O Neely M Panton, and M Stella Atkins. Workload assessment of surgeons: correlation between nasa tlx and blinks. *Surgical endoscopy*, 26(10):2746–2750, 2012.

[46] Ferdinand Rudolf Hendrikus Zijlstra. Efficiency in work behaviour: A design approach for modern tools. 1995.

Appendix A

Gaussian Process Algorithm Validation

In order to ensure the GP algorithm performs correctly, the algorithm was validated using a data set from an example in *GPFlow* Python library shown in Figure A.1.

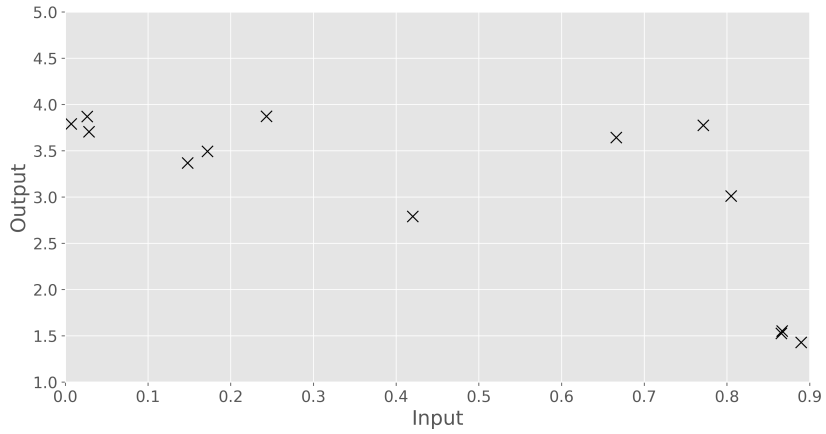


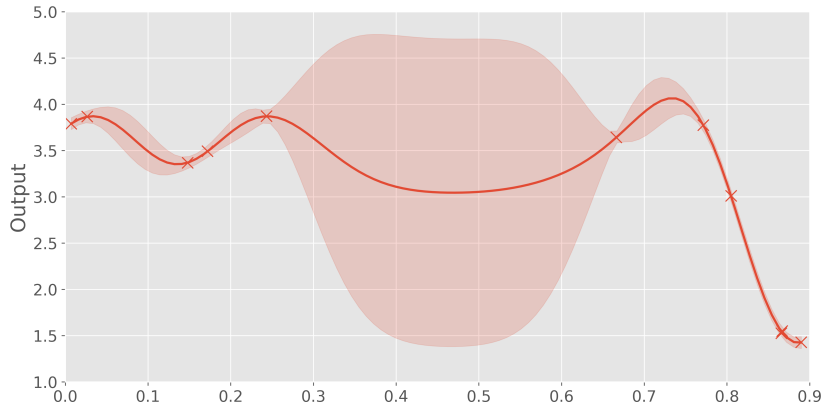
Figure A.1: An arbitrary 2-D data set used for validating the GP algorithm.

Using the GPy python library, the algorithm creates an initial GP model with a SE kernel based on an initial desired number of data points (Figure A.2a). The

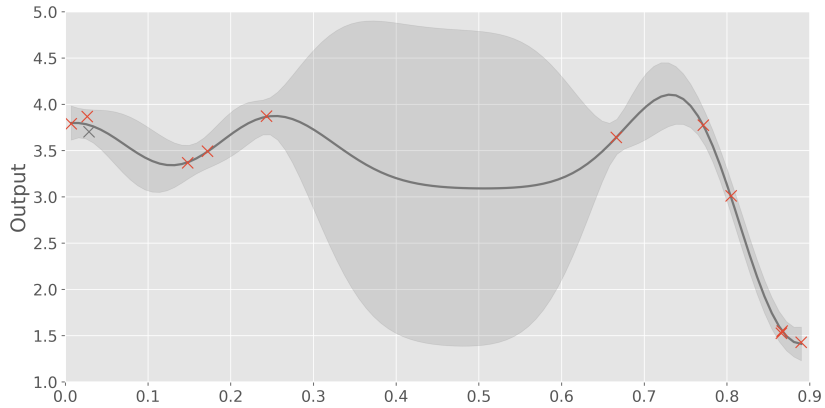
SE kernel enables ARD, which implicitly determines the relevance of each input by creating a length-scale parameter for each input. While ARD does not affect single input modeling, enabling ARD allows further sifting for informative data. The GP fits a mean function to the data and reports associated variances along that function. Figure A.2 details the conversion of variances to confidence intervals to show low uncertainty near data points and high uncertainty in between data. As the GP receives new data, the model updates and optimizes hyperparameters (Figure A.2b-c).

By adding a third dimension, the number of data points considered, we can view the progression of the model as a surface (Figure A.3). Each updated model appends to the plot as a line and then interpolated to extend the surface along the number of data points axis.

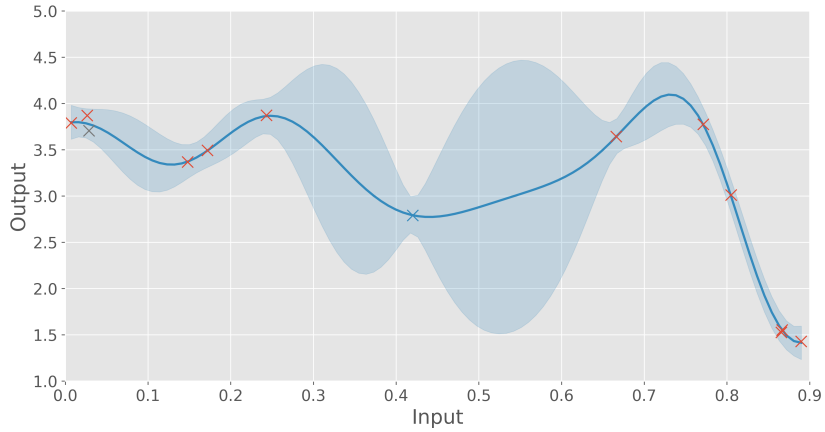
The algorithm produces an evolution of the model based on the validation data set. While the GP model presented in this work excludes time, using the number of data as an axis served the same purpose. To verify GP model reliability, the algorithm calculated performance surfaces for each performance metric based on the first principal component (Figure 5.2). Knowing the GP algorithm produces trustworthy results with the validation data set and simplified physiological data, GP we have confidence in the multivariate GP model performance predictions explained herein.



(a)



(b)



(c)

Figure A.2: Simulating online modeling, the GP updates with each new data point. Model (a) includes 11 data points with (b) and (c) adding one more data point for each. The shaded regions represent a 95% confidence interval. The highest uncertainty sits between large gaps of data, but when new data fills the gap, the uncertainty drops.

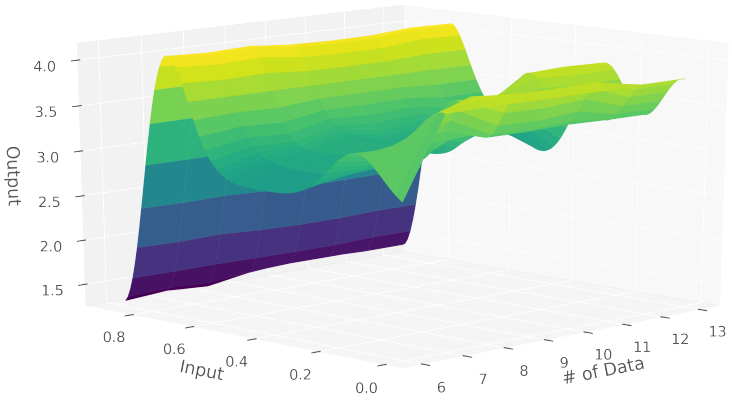


Figure A.3: GP model evolution with each additional data point. Starting with a model of 6 data points, the model successively updates with each new data point. Using the number of data points as one of the axes, the models create a surface. For the presented GP algorithm, time takes the place of number of data points.