**Time-Series Regression for Users of Redback Operations Sport Applications: Overview and Problems**

**Overview:**

Given Redback Operations interest in improving user performance in areas of exercise and wellbeing, one area that could be a point of focus for the Data Science & Analysis team is regression techniques. Regression techniques are powerful algorithmic tools designed to predict continuous or floating-point numbers (Muller & Guido 2016). Our goal at Redback Operations is to improve user performance based on time-series data in cycling, running, or walking simulators produced by our game development team. Leeuw et al (2020) used time-series data from professional cyclists to provide an accurate prediction of the next value in a sequence of races – and aggregate those predictions to give an overall sense of whether a cyclist was losing time or gaining time over the course of a race season.

I believe that such predictions can be useful not only to users – who may like to identify whether they are improving over time or not, but also to Redback Operations itself. A fundamental goal of Redback Operations is to provide real-time and asynchronous feedback to our users via data from users that has been processed and analysed by algorithms we have designed. A time-series regression analysis of the times that our users are completing certain exercise tasks may help us calibrate our own automated decision-making processes to better serve the user.

As stated above, regression technique are powerful tools that allow us to predict certain outcomes given inputs. However, much of the data that we have access to in exercise and well-being initiatives comes in the form of time-series, where the observations we encounter are dependent, and as Box (1991) succinctly states: “where the nature of this dependence is of interest in itself” (Box 1991: ix). However, as Leeuw et al (2020) rightly point out, the focus of time-series analysis has largely been about forecasting rather than regression tasks. We engage here in regression rather than forecasting because our aim “is to determine a target value for a yet to be predicted time series based on a collection of historical time series and target values” (Leeuw et al 2020: 690). The difference between forecasting and regression lies in the location of the value being predicted. Whereas in forecasting one is interested in predicting the next value in a sequence of values, “in time-series regression we are primarily interested in learning how to aggregate a time series into a single number” (Leeuw et al 2020) which would allow us to indicate to the user whether they are losing or gaining time over a period; in other words, whether the user is improving or not based on their exercise regime

As stated by Leeuw et al (2020) the time-series regression problem can be seen as the regression counterpart to the time-series classification problem. Instead of predicting a certain label however, the regression problem seeks to predict a value for an unseen time series. In the scenario painted by Redback Operations, a user would in using our products would automatically submit the time taken to complete certain objectives in a simulated course (such as passing marker e at time t); along with other data collected via IoT sensors and the in-game simulation. A simple yet effective comparison between users is to show the times it takes for any individual user to reach marker e during the simulation. However, for more comprehensive feedback, we should be able to predict the time that a user takes to reach marker e based on previous attempts at the simulated course. This is where I think time-series regression can be a powerful tool. By being able to predict the time that a specific user takes to reach a point in the simulated course based on their previous attempts at the simulated course, we can then begin to pinpoint training and improvement points to help them become the best person they can be.

**Goal:** The primary goal of this research is to get to a stage where we can produce an algorithm which can predict the performance of a user in any simulated course solely based on the characteristics of the rider and the simulated course.

**Problem 1**: What characteristics of the user and of the simulated course will be most beneficial to producing an effective, accurate output?

This problem can further be split into a series of related problems around feature selection:

* What length of time-series data is required?
* What aspects of time-series are relevant to the prediction?
* What aspects of the user’s characteristics will be relevant to prediction?

For example: is it possible to use multiple simulated courses of different lengths, with varying terrains to predict the performance of a user on a yet un-trained simulated course?

Will the target value depend on continuous, aggregated features such as the total amount of climbs within a simulated course, or the total usage of brake-age in a course to come to an accurate total, or will the target value depend on more discrete forms of those features?

How will the general characteristics of the user such as height, weight, initial bike/walking speed, be introduced into a purported time-series analysis?

Several solutions: LASSO (Tibshirani 1996) , Hierarchical Representation using NMF (Song et al 2013), Feature-based representations (Fulcher & Jones 2014)

**Lasso (Least absolute shrinkage and selection operator):** LASSO restricts coefficients to be close to zero. In this form of regularization, the coefficients of some features are reduced entirely to zero – and thereby ignored through a system of automatic feature selection. LASSO might be a useful choice given we are unable to identify what features might be most useful for our predictive algorithm

**Hierarchical Representation using NMF**: Introduced by (Song et al) in 2013, the proposed network discovers feature hierarchies present in complex data and demonstrates them in intuitively understandable manner by learning feature relationships among the layers in non-negative approach. By simple addition and accumulation of features, we are able to understand the data structure and construct a hierarchy based on the information learned by the network (Song et al 2013: 474). A potential problem with this, not mentioned by Song, is the difficulty in understanding how features are added, combined or eliminated in the process. We do not know how the hierarchical learning model is making judgements about certain features – and do not know whether it is eliminating features that may actually have a correlation to the time in which users complete the simulated course.

**Feature-based Representations:** Fulcher & Jones (2014) explore whether feature-based representation for feature selection can be used on more long term time-series, such as they used in short pattern-like time series. “The process is completely data-driven and does not require any knowledge of the dynamical mechanisms underlying the time series or how they were measured.” (Fulcher & Jones: 3). Problems, computational expensive and data-rich. Perhaps too general, as they are just measuring what time of parameters are most useful in time-series analysis over a whole array of datasets which are not domain specific.

**Problem 2:** To be able to test the hypotheses raised in the above discussion regarding the algorithm and feature selection, the team needs to be able to use specific data that contains several weeks or months of time-series data related to one user. Leeuw et al (2020) were able to work in conjunction with a specific professional cycling team for their data. This may take some time – and so for now the goal of producing such an algorithm may have to be put on the backburner until more sophisticated data is made available to Redback Operations.