UCL CDT DIS Note

UCLCDTDIS-2019-XX

20th April 2020

Results from TBR Group Project

Petr Mánek[a] and Graham Van Goffrier[a]

[a]University College London

The abstract of my report.

Contents

1 Introduction

The analysis of massive datasets has become a necessary component of virtually all technical fields, as well as the social and humanistic sciences, in recent years. Given that rapid improvements in sensing and processing hardware have gone hand in hand with the data explosion, it is unsurprising that software for the generation and interpretation of this data has also attained a new frontier in complexity. In particular, simulation procedures such as Monte Carlo (MC) event generation can perform physics predictions even for theoretical regimes which are not analytically soluble. The bottleneck for such procedures, as is often the case, lies in the computational time and power which they necessitate.

Surrogate models, or metamodels, can resolve this limitation by replacing a resource-expensive procedure with a much cheaper approximation [1]. They are especially useful in applications where numerous evaluations of an expensive procedure are required over the same or similar domains, e.g. in the parameter optimisation of a theoretical model. The term "metamodel" proves especially meaningful in this case, when the surrogate model approximates a computational process which is itself a model for a (perhaps unknown) physical process [2]. There exists a spectrum between "physical" surrogates which are constructed with some contextual knowledge in hand, and "empirical" surrogates which are derived purely from the underlying expensive model.

In this internship project, in coordination with the UK Atomic Energy Authority (UKAEA) and Culham Centre for Fusion Energy (CCFE), we sought to develop a surrogate model for the tritium breeding ratio (TBR) in a Tokamak nuclear fusion reactor. Our expensive model was a MC-based neutronics simulation [3], itself a spherical approximation of the Joint European Torus (JET) at CCFE, which returns a prediction of the TBR for a given reactor configuration. We took an empirical approach to the construction of this surrogate, and no results described here are explicitly dependent on prior physics knowledge.

For the remainder of Section 1, we will define the TBR and set the context of this work within the goals of the UKAEA and CCFE. In Section 2 we will describe our datasets generated from the expensive model for training and validation purposes, and the dimensionality reduction methods employed to develop our understanding of the parameter domain. In Section 3 we will present our methodologies for the comparison testing of a wide variety of surrogate modelling techniques, as well as a novel adaptive sampling procedure suited to this application. After delivering the results of these approaches in Section 4, we will give our final conclusions and recommendations for further work.

1.1 Problem Description

Nuclear fusion technology relies on the production and containment of an extremely hot and dense plasma. In this environment, by design similar to that of a star, hydrogen atoms attain energies sufficient to overcome their usual electrostatic repulsion and fuse to form helium [4]. Early prototype reactors made use of the deuterium (^2H) isotope of hydrogen in order to achieve fusion under more accessible conditions, but achieved limited success. The current frontier generation of fusion reactors, such as JET and the under-construction International Thermonuclear

Experimental Reactor (ITER), make use of tritium (^3H) fuel for further efficiency gain. Experimentation at JET dating back to 1997 [5] has made significant headway in validating deuterium-tritium (D-T) operations and constraining the technology which will be employed in ITER in a scaled up form.

However, tritium is much less readily available as a fuel source than deuterium. While at least one deuterium atom occurs for every 5000 molecules of naturally-sourced water, and may be easily distilled, tritium is extremely rare in nature. It may be produced indirectly through irradiation of heavy water (deuterium oxide) during nuclear fission, but only at very low rates which could never sustain industrial-scale fusion power.

Instead, modern D-T reactors rely on tritium breeding blankets, specialised layers of material which partially line the reactor and produce tritium upon neutron bombardment, e.g. by

$$ {}^1_0n + {}^6_3Li \rightarrow {}^3_1T + {}^4_2He \tag{1} $$

$$ {}^1_0n + {}^7_3Li \rightarrow {}^3_1T + {}^4_2He + {}^1_0n \tag{2} $$

where T represents tritium and ^7Li, ^6Li are the more and less frequently occurring isotopes of lithium, respectively. ^6Li has the greatest tritium breeding cross-section of all tested isotopes [4], but due to magnetohydrodynamic instability of liquid lithium in the reactor environment, a variety of solid lithium compounds are preferred.

The TBR is defined as the ratio between tritium generation in the breeding blanket per unit time and tritium fuel consumption in the reactor. The MC neutronics simulations previously mentioned therefore must account for both the internal plasma dynamics of the fusion reactor and the resultant interactions of neutrons with breeding blanket materials. Neutron paths are traced through a CAD model (e.g. Figure 1) of a reactor with modifiable geometry.

The input parameters of the computationally-expensive TBR model therefore fall into two classes. Continuous parameters, including material thicknesses and packing ratios, describe the geometry of a given reactor configuration. Discrete categorical parameters further specify all relevant material sections, including coolants, armours, and neutron multipliers. One notable exception is the enrichment ratio, a continuous parameter denoting the presence of ^6Li. Our challenge, put simply, was to produce a TBR function which takes these same input parameters and approximates the MC TBR model with the greatest achievable accuracy.

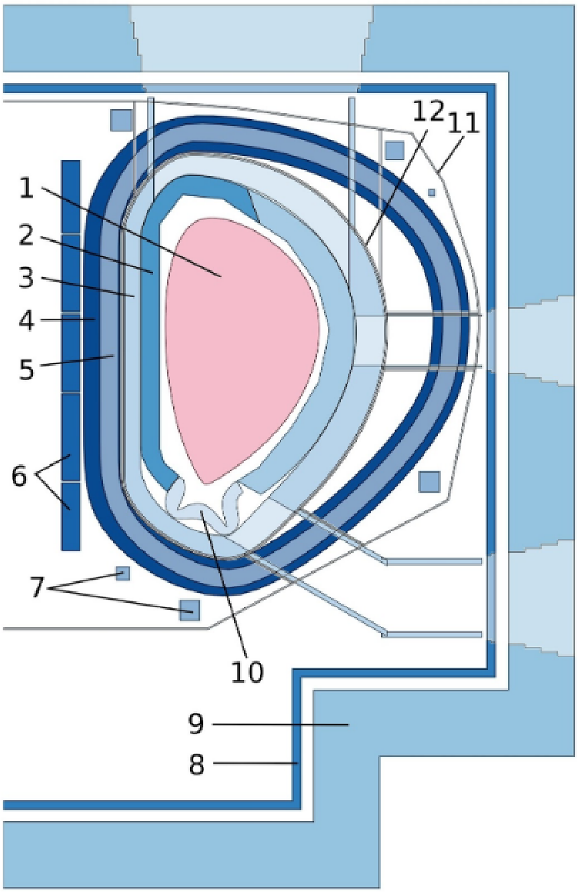

Figure 1: Typical single-null reactor configuration as specified by BLUEPRINT [6]: 1 — plasma, 2 — breeding blankets

2 Data Preprocessing

Data

2.1 Data Description and Initial Sampling

2.2 Dimensionality Reduction

Model training over high-dimensional parameter spaces may be facilitated by carefully reducing the number of variables used to describe the space. For many applications, feature selection strategies succeed in identifying a sufficiently representative subset of the original input variables; however, all given variables were assumed to be physically relevant to the MC TBR model. Feature extraction methods, on the other hand, aim to identify a transformation from the parameter space which decreases dimensionality; even if no individual parameter is separable from the space, some linear combinations of parameters or nonlinear functions of parameters may be.

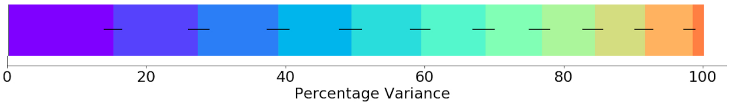

Figure 2: Cumulative variance for optimal features identified by PCA

2.2.1 Principal Component Analysis

To pursue linear feature extraction, principal component analysis (PCA) was performed via scikit-learn on a set of 300,000 uniform samples of the MC TBR model.

Figure 2 shows the resultant cumulative variance of the feature vectors identified by PCA. The similar share of variance among all eleven features demonstrated irreducibility of the TBR model by linear methods.

2.2.2 Variogram Computations

Kriging, or Gaussian process regression, is a geostatistical surrogate modelling technique which relies on correlation functions over distance (lag) in the paramater space [2]. Although kriging performed poorly for our use case due to high dimensionality, these correlation measures gave insight into similarities between discrete-parameter slices of the data.

Figure 3 shows the Matheron semivariance [3] for three discrete slices with coolant material varied, but all other discrete parameters fixed. Fits [4] to the Matérn covariance model confirmed numerically that the coolant material is the discrete parameter with the greatest distinguishability in the MC TBR model.

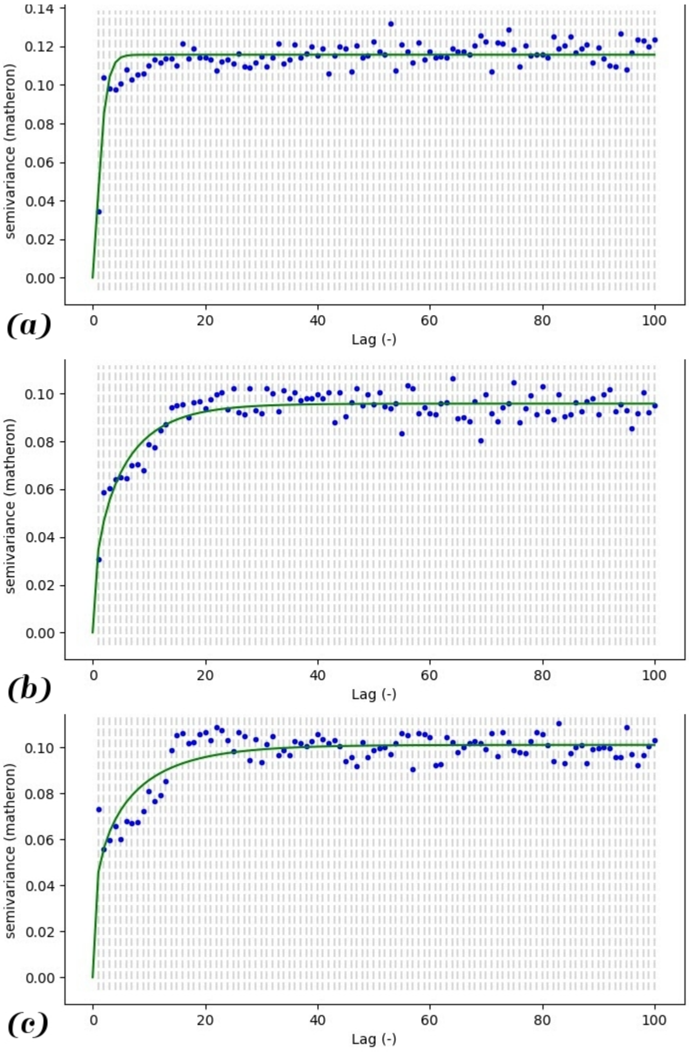

Figure 3: Semivariograms for MC TBR data with coolant materials: (a) He, (b) H_2O, (c) D_2O

** Make sure discrete slices are mentioned before this **

2.2.3 Autoencoders

Question: is it accurate to say this represents nonlinear feature extraction? re my intro paragraph

3 Methodology

Assuming that input has been appropriately treated to eliminate redundant features, we may turn to characterise proposed surrogate models and the criteria used for their evaluation. The task all presented surrogates strive to solve can be formulated using the language of conventional regression problems. In the scope of this work, we explore various possible choices available to us in the scheme of supervised and unsupervised learning.

Labeling the expensive Monte Carlo simulation $f(x)$, a surrogate is a mapping $\tilde{f}(x)$ that yields similar images as $f(x)$. In other words, $f(x)$ and $\tilde{f}(x)$ minimise a selected similarity metric. Furthermore, in order to be considered *viable*, surrogates are required to achieve expected evaluation time that does not exceed the expected evaluation time of $f(x)$.

In the supervised learning setting, we first gather a sufficiently large training set of samples $\mathcal{T} = \left\{\left(x^{(i)}, f\left(x^{(i)}\right)\right)\right\}_{i=1}^{N}$ to describe the behaviour of $f(x)$ across its domain. Depending on specific model class and appropriate choice of its hyperparameters, surrogate models $\tilde{f}(x)$ are trained to minimise empirical risk with respect to $\mathcal{T}$ and a model-specific loss function $\mathcal{L}$, where empirical risk is defined as

$$R_{\text{emp.}}(\tilde{f} \mid \mathcal{T}, \mathcal{L}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\tilde{f}(x^{(i)}), f(x^{(i)})\right). \tag{3}$$

The unsupervised setting can be viewed as an extension of this method. Rather than fixing the training set $\mathcal{T}$ for the entire duration of training, multiple sets $\{\mathcal{T}_k\}_{k=0}^{K}$ are used, such that $\mathcal{T}_{k-1} \subset \mathcal{T}_k$ for all k. The first set $\mathcal{T}_0$ is initialised randomly to provide a *burn-in*, and is repeatedly extended in epochs. In each epoch, a new surrogate is trained on $\mathcal{T}_k$ using the supervised learning procedure, and evaluated. This permits the learning algorithm to motivate the choice of new points following the evaluation of surrogates trained thus far by appropriately biasing a proposal distribution, in order to focus on improvement of complex regions within the domain.

3.1 Metrics

Aiming to provide objective comparison of a diverse set of surrogate model classes, we define a multitude of metrics to be tracked during experiments. Following the motivation of this work, we are primarily interested in two properties of surrogates: (i) their capability to approximate the expensive model well and (ii) their time of evaluation. We strive to maximise the former while minimising the latter.

To prevent undesirable bias in results due to training set selection, both properties are collected using k-fold cross-validation with a standard choice of $k = 5$. Herein, a sample set is subdivided into 5 disjoint folds which are repeatedly interpreted as training and testing sets, maintaining a constant ratio of samples. In each such interpretation experiments are repeated, and the overall value of each metric of interest is given by the mean across all folds.

The remainder of this section provides exhaustive list and description of regression performance and evaluation time metrics recorded in the experiments.

3.1.1 Regression Performance Metrics

TODO: provide formal definition of each of the following metrics

Mean absolute error

R^2 ratio

Adjusted R^2 ratio

Standard error of regression

3.1.2 Evaluation Time Metrics

TODO: provide formal definition of each of the following metrics

Training time per sample

Prediction time per sample

3.2 Model Comparisons

3.3 Adaptive Sampling

4 Results

Results

4.1 Results of Model Comparisons

4.2 Results of Adaptive Sampling

5 Conclusion

Conclusion