Proceedings of the International Conference on Applied Mechanical Engineering Research (IC-AMER2019)
NIT Warangal, India – May 02-05, 2019
Paper No. XXX (The number assigned when the paper is accepted.)

Identifying empirically important variables in IC engine operation through redundancy analysis

Satishchandra Salam[a]*, Tikendra Nath Verma[a]
[a]*Department of Mechanical Engineering, National Institute of Technology Manipur, Imphal-795004, India*

*Corresponding author Email: satisji@gmail.com

Computational studies incur engineering costs. While direct numerical simulations can provide detailed solutions, they cannot deliver quick and convenient solutions which are pragmatic for industry applications such as fault detection and diagnosis. But in the study of internal combustion engines, there are no such unified models that can completely capture the engine operation and hence, computational methods still are of great value. In this pursuit, an attempt had been made in this study to evaluate the empirical redundancy amongst the engine variables. Through the methodology presented in this study, those empirically important variables can be identified and they can be used to develop empirically reduced models for its possible employment in further computational studies.

***Keywords*:** correlation matrix; redundancy analysis; empirical modelling; IC engine.

1. Introduction

Even almost after 150 years of its introduction, the pursuit for engineering solutions of internal combustion engine (ICE) still continues. Essentially a heat engine that converts the chemical energy of the fuel to mechanical energy, it fundamentally involves controlled combustion of fuel inside the combustion chamber. Therefore, the process is influenced even by small length and time scales. And as ICE research evolves, the paradigm has shifted from the fundamental analytical modelling to phenomenological modelling, and in the last few decades to computational modelling (Yu et al., 2011).

The success with computational modelling has been largely achieved with improved computational fluid dynamics (CFD) based methods which can deliver detailed numerical solutions (DNS). By decomposing large real-life systems into smaller subsystems, analytical models are implemented at the sub-system level to evaluate the phenomenon associated with the whole system. This has delivered solutions which are of extreme value, and are even sometimes left as the only feasible method.

But despite how CFD can deliver those DNS, it incurs large engineering resources including time and capital. There is a significant trade-off between the cost of method and the accuracy of solution it can deliver (Fig. 1). For instance, when all the combinatorics of all operating conditions are to be considered for the full-scale simulation of studying technical feasibility of various biofuels as alternative fuel, CFD based methods racks up large computational cost which are not feasible for industry applications. This (therefore) calls for alternative methods which can deliver quick and convenient industry feasible solutions for diagnosis and fault detection.

One such possible alternative method could be delivered by what has been popularly known as ‘data reduction’ in the emerging applications of data science (Bevington et al., 1993, Hinton and Ruslan, 2006). In this mathematically black-box modelling method, the objective is to identify the empirical pattern embedded in the observation regardless of what mechanistic explanation the model can offer. Mathematically, when the number of variables is greater than the number of

equations, multiple sets of solutions can exist. These various solutions are superfluous and are alternative to one another. While each solution is unique, the involved variables can accommodate variability amongst a set of the involved variables. This redundancy when perceived from empirical perspective for the purpose of black-box modelling is hereby referred to as empirical redundancy. This can be achieved by sequestrating redundancy from the dataset whose detailed methodology in the case of ICE operation will be discussed in the following sections.

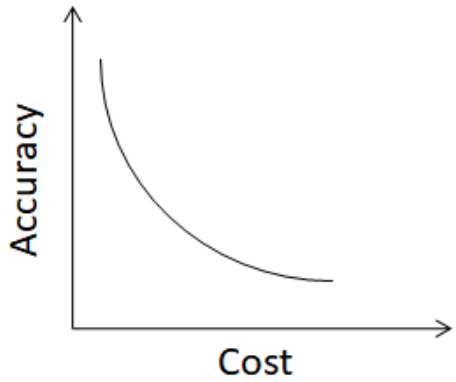

Fig. 1. Illustration of cost-accuracy trade-off.

2. Methods

Salam and Verma, 2019 and Pankaj et al., 2019 (unpublished observations) have reported empirical redundancy in the engine responses. The correlation matrices reported in those studies suggested methods to harness the redundancy in the system. Here, using the correlation matrix reported in Salam and Verma, 2019 as sample correlation matrix, the following methodology is proposed to evaluate those reported redundancy. The study used inputs of loading, blending and fuel injection pressure to characterise for performance with specific fuel consumption (SFC), brake thermal efficiency (BTE), indicated efficiency (IE) and scavenging efficiency (SE), combustion with exhaust gas temperature (EGT), cylinder peak pressure (CPP), cylinder peak temperature (CPT), maximum rate of pressure rise (MRPR), outer mean diameter of injected droplet (Dout) and ignition delay (ID), and emission with Hartridge smoke unit (HSU), smoke, specific particulate matter (SPM), carbon dioxide (CO2), oxides of nitrogen (NOx), summary of emission (SoE) and nitrous oxide (NO2).

2. 1. Correlation matrix

Correlation indicates how strongly a variable is related to another variable. Of different indicators of correlation, Pearson correlation coefficient (ρ) is one such indicator which can quantify the monotonic dependencies among the set of variables. Mathematically, for two matrixes X(a) and Y(b) with means $\overline{X_a}$ and $\overline{Y_b}$ respectively, it is defined as (1):

$$\rho(a,b) = \frac{\Sigma(X_{a,i}-\overline{X_a})(Y_{b,i}-\overline{Y_b})}{\left\{\Sigma(X_{a,i}-\overline{X_a})^2\Sigma(Y_{b,j}-\overline{Y_b})^2\right\}^{1/2}} \quad (1)$$

For the empirical analysis of the engine variables, this was performed using the 'corr' function available in MATLAB 2016a. These correlation coefficients across all the variables were presented in Fig. 2. It can be noted that the diagonal cells had a value of unity since they represent self-correlation.

2. 2. Cumulative Histogram

A histogram was drawn for the correlation coefficients with a bin size of 20(=√20x20, Shimazaki and Shinomoto, 2007). Since we are interested only in those strongly correlated

variables regardless of the directionality, the histogram was computed for absolute of the correlation coefficients. And as it was evident from Fig. 3, the correlation coefficients were sparsely distributed over [0,1] with significant counts towards correlation coefficient of 1. This was indicative of how several variables were strongly related to each other empirically. Kindly note that it had not yet been discussed about how these empirical dependencies actually manifest into functional mechanistic dependencies.

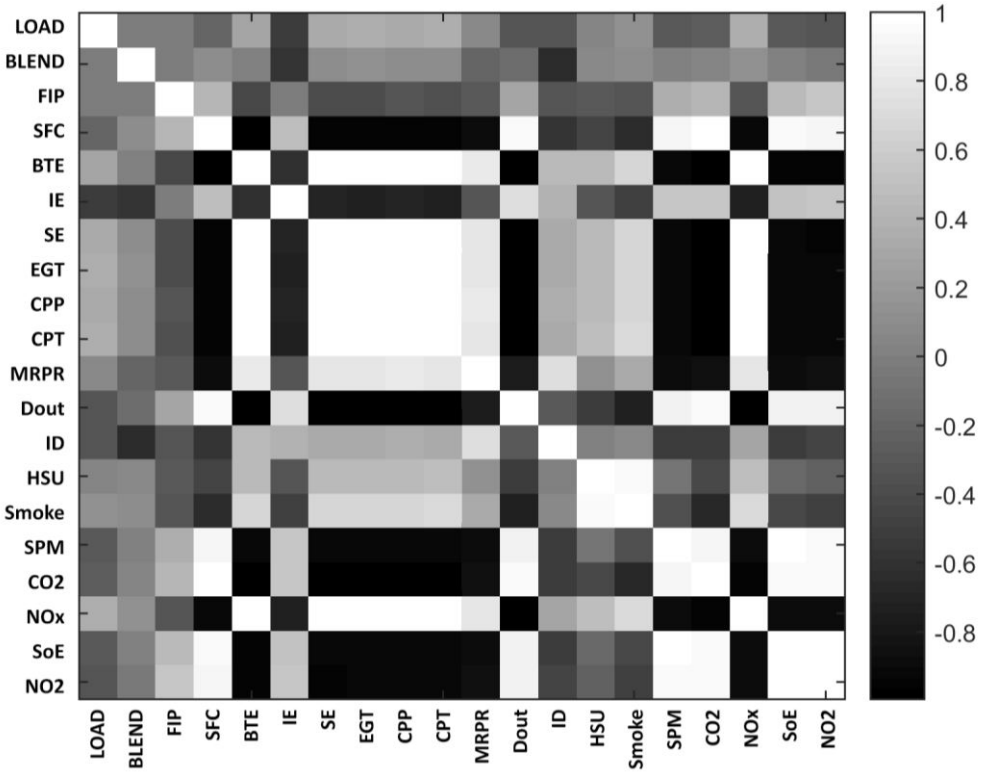

Fig. 2. Correlation matrix of the engine variables.

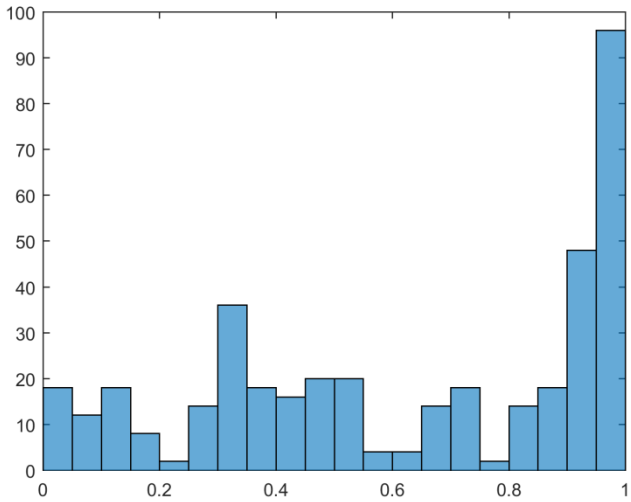

Fig. 3. Histogram of absolute of Pearson correlation coefficients.

2. 3. Representation score

To quantify how efficient a variable is in representing a set of redundant variables, an index hereby referred to as representation score was introduced (2):

$$Representation\ score\ of\ i^{th}\ variable = \frac{1}{n}\sum(correlation\ coefficient\ with\ i^{th}\ variable) \quad (2)$$

By averaging the correlation coefficients for a variable with all other variables, representation score provided a basis for comparing how a variable (as compared with other variables) can effectively represent other variables. Larger the representation score is, more efficient it is in representing the whole set of variables. Thus, variables with higher representation score are the more suitable variables to empirically represent the whole system. For this set of variables, the sorted variables in order of decreasing representation score were presented in Fig. 4.

2. 4. Thresholding

To decide on exactly how many variables should be picked as representative variables to substitute for the whole system, this thresholding section is added only to assist the designer in the decision making process. While there is the obvious upper limit of 100% reconstruction accuracy when all the variables are chosen, it entirely rests on the designer to select a lesser number of variables under the constraint of cost-accuracy trade-off as discussed earlier.

An alternative here is to maximise the ratio of accuracy to number of variables i.e., to get the maximum accuracy out of least number of variables. This can be mathematically achieved by taking the first derivative of profile of the representation score ranking as in Fig. 4. A second rate change will help decide the local extremas of the profile (i.e., either local maxima or minima).

3. Results and discussion

Fig. 4 shows all the engine associated variables sorted in order of empirical importance. The superficial interpretation would be that variables having highest representation score would be the most important variables from empirical perspective. These variables have the highest linear dependencies across all other variables as well as amongst all other variables. Consequently, they are the best candidates to empirically represent other variables.

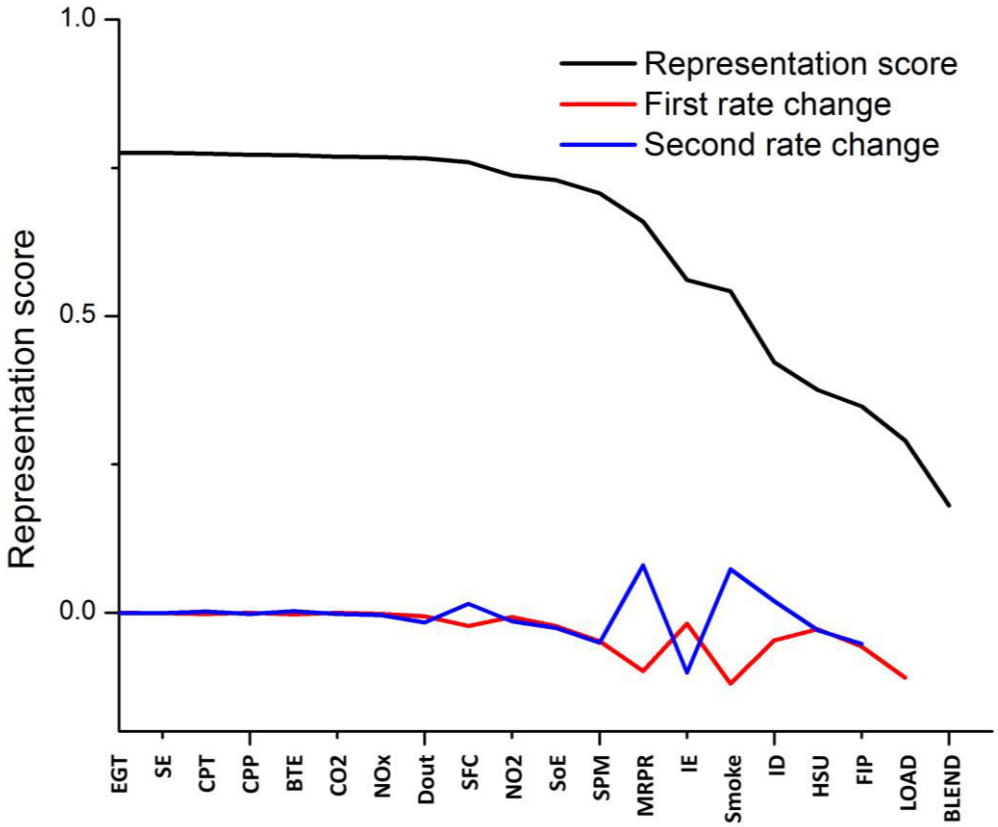

Fig. 4. Engine variables sorted by representation scores.

As discussed on thresholding, the reconstruction accuracy will be affected by the number of variables chosen for data reconstruction. And as it is unlikely for the reconstruction accuracy to be directly proportional to the number of variables, the 'best' number of choice under the constraint of cost-accuracy trade-off can be decided by picking the threshold value which can be decided by the designer.

In closure, while this study analysed the empirical dependencies, incorporating the functional mechanistic relations amongst the variables was out of scope of the study. It had not commented on how these empirically identified 'important' variables are influencing the actual ICE operation. This lack of explanation was inherently because of why 'correlation is not causality' (Harford, 2014), and on how black box models cannot explain the cause and effect relationships. Therefore, enough emphasis should be given to dataset specific interpretation with domain knowledge for the proper implementation of the presented methodology.

4. Conclusion

The study had presented a systematic methodology to empirically reduce a set of variables associated with ICE operation. After employing Pearson correlation coefficient to quantify the monotonic dependencies among the engine variables, an index called representation score had been used to compare how efficiently a variable can represent the whole system. Such models will help derive quick and convenient solutions for a large class of industry applications.

References

1. Bevington, Philip R., et al. "Data reduction and error analysis for the physical sciences." *Computers in Physics* 7.4 (1993): 415-416.

2. Harford, Tim. "Big data: A big mistake?" *Significance* 11.5 (2014). 14-19.

3. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006). 504-507.

4. Pankaj Shivastava, Salam, Satishchandra, and Tikendra Nath Verma. "Experimental and empirical study of CI engine operating with Lal ambari biodiesel." submitted to *Energy Conversion and Management* (2019). (unpublished observations)

5. Salam, Satishchandra, and Tikendra Nath Verma. "Appending empirical modelling to numerical solution for behaviour characterisation of microalgae biodiesel." *Energy Conversion and Management* 180 (2019): 496-510.

6. Shi, Yu, Hai-Wen Ge, and Rolf D. Reitz. "Computatioanl optimisation of internal combustion engines." *Springer Science and Business Media* (2011).

7. Shimazaki, Hideaki, and Shigeru Shinomoto. "A method for selecting the bin size of a time histogram." *Neural computation* 19.6 (2007). 1503-1527.