

Multiple Kernel Kriging Methods for Aerospace Applications

Kehinde Sikirulai Oyetunde [20548072], Sathi Sarveswara [20647048]

Abstract

Computational fluid dynamics is used in aerospace industries to simulate computational experiments. The computations involve with CFD usually requires several iterations and could take days to solve depending on the size of the model. The heavy computational power required makes it difficult to solve certain design optimization problems. Kriging (Gaussian processes) are introduced to build high performance surrogates for such experiments using limited data. This report focuses on the development and use of machine learning techniques (Kriging) for small datasets. In specific, we used various kernels within Kriging formulations. Also, three multiple kernel methods are developed and benchmarked with four single kernels. A 8D and 16D CFD data which have the drag coefficient as its quantity of interest (QoI) is used. The composite kernel learning multiple kernel method performs best with lesser model error and computational times.

1 Introduction

The work done in this project is motivated by the heavy computational resource required for CFD simulations in the aerospace industries. Also, in part motivated by the need for machine learning techniques for applications with small datasets having a little as 1000 points. We developed Kriging models which are a variant of Gaussian process regression (Gpr) but with no noise. Hence, Kriging models are best suited for deterministic data obtainable in CFD simulations. CFD simulations are considered deterministic because same set of input will give same output. Our Kriging methods models the output as a realization of a Gaussian distribution. We used the Gaussian, exponential, Matérn 3/2, Matérn 5/2 kernels in building our models. Three multiple kernels methods are developed using the four single kernels to improve the performance of the Kriging models as well as reduce the bias from the individual kernels. Lastly, we compared the performance of the seven models (single and multiple kernel methods) on the two available datasets.

2 Methods

Kriging is used for approximating computationally expensive blackbox functions. The relationship between the input, \mathbf{x} , and the output, \mathbf{y} of a blackbox function in Kriging, can be expressed as:

$$Y(x) = \mu(x) + Z(x), \quad (1)$$

where $\mu(x)$ is the deterministic function approximating the mean trend of the output and $Z(x)$ is the stochastic process which deviates from this trend. $Z(x)$ has a zero mean and the covariance $\text{Cov}[Z(\mathbf{x}), Z(\mathbf{x}')] = \sigma^2 \mathbf{R}(\mathbf{x}, \mathbf{x}')$. σ^2 is the process variance and $\mathbf{R}(\cdot, \cdot)$ is the correlation matrix. The correlation between the i -th and j -th samples can then be expressed as:

$$R_{ij}(\mathbf{h}, \boldsymbol{\theta}) = \prod_{k=1}^{N_d} R(h_{ij}^{(k)}, \theta^{(k)}). \quad (2)$$

The vector of correlation parameters is denoted as $\boldsymbol{\theta} = \{\theta^{(k)}\}$, $k = 1, \dots, N_d$. The value $h_{ij}^{(k)}$ is the distance between two points in the k^{th} dimension, $|x_i^{(k)} - x_j^{(k)}|$. For a simple kriging model with a constant trend, the mean and the variance of the prediction at a point \mathbf{x} is expressed as;

$$m(\mathbf{x}) = \mu + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\mu) \quad (3)$$

$$s^2(\mathbf{x}) = \hat{\sigma}^2 (1 - \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x})) \quad (4)$$

2.1 Kernels

Kernels are used to capture the non-linear relationship that exist between the input \mathbf{x} and output, \mathbf{y} . In this report, four kernels: the Gaussian, Exponential, Matérn 3/2 and Matérn 5/2 kernels are used for performance comparison and also to develop the various models.

- Gaussian

$$R(\mathbf{h}, \boldsymbol{\theta}) = \exp -\frac{1}{2} \left(\frac{\mathbf{h}}{\boldsymbol{\theta}} \right)^2 \quad (5)$$

- Exponential

$$R(\mathbf{h}, \boldsymbol{\theta}) = \exp -\frac{1}{2} \left(\frac{\mathbf{h}}{\boldsymbol{\theta}} \right) \quad (6)$$

- Matern 3/2

$$R(\mathbf{h}, \boldsymbol{\theta}) = \left(1 + \frac{\sqrt{3}|\mathbf{h}|}{\boldsymbol{\theta}} \right) \exp \left(-\frac{\sqrt{3}|\mathbf{h}|}{\boldsymbol{\theta}} \right) \quad (7)$$

- Matern 5/2

$$R(\mathbf{h}, \boldsymbol{\theta}) = \left(1 + \frac{\sqrt{5}|\mathbf{h}|}{\boldsymbol{\theta}} + \frac{5\mathbf{h}^2}{3\boldsymbol{\theta}^2} \right) \exp \left(-\frac{\sqrt{5}|\mathbf{h}|}{\boldsymbol{\theta}} \right) \quad (8)$$

where h is a function of distance $|x - x'|$ and $\boldsymbol{\theta}$ is the length scale

2.2 Multiple kernel methods

This generally refers to methods which uses more than one kernel or several instance of a kernel (with different parameters). We developed three multiple kernel methods; ensemble method, composite kernel learning (CKL), multidimensional composite kernel learning (MCKL) methods. The ensemble method [3, 4] which can either be global or local uses weights to combine different models. We developed the ensemble method which uses corrected-Akaike information criterion (AICc) [1] for global weight computations. The correction introduced takes into account the factor of sample size and increase the penalty for model complexity with small data sets. The ensemble method help to reduce error and bias in the model. Another method which we developed is the Composite Kernel Learning (CKL) method [2]. It involves the construction of new kernels by combinations of existing kernels. We propose the multidimensional composite kernel method that derives unique weight vectors for each variable. It is aimed at greatly improving the robustness of Kriging models. The technique aims at improving the approximation quality of models by avoiding the misspecification of kernels.

- **Ensemble method:** The predictions of each models are globally combined using equation 9

$$\hat{y}_{ens}(x) = \sum_{i=1}^K w_i(x) \hat{y}_i(x) \quad (9)$$

The weight is computed;

$$w_i = \frac{\exp(-0.5\Delta AIC_i)}{\sum_{j=1}^K \exp(-0.5\Delta AIC_j)} \quad (10)$$

where,

$$AIC = -2\ln(L) + 2N_f \quad (11)$$

$$AIC_c = AIC + \frac{2N_f^2 + 2N_f}{n - N_f - 1} \quad (12)$$

$$\Delta AIC_i = AIC_i - AIC_{min} \quad (13)$$

$$N_f = \text{len}(\theta) + 2 \quad (14)$$

- **Composite Kernel Learning (CKL):** The weighted sum formulation of combined kernels as follows:

$$R_{CKL}(h) = \sum_{i=1}^K w_i(x) R_i(h) \quad (15)$$

where $\mathbf{w} = w_1, w_2, \dots, w_K^T$ is the non-negative weight vector that should satisfies the equality constraint $\sum_{i=1}^K w_i = 1$.

- **Multidimensional Composite Kernel Learning (MCKL)**

The kernels for the MCKL formulations are combined using equation 16

$$R_{ij}(\theta, h) = \prod_{d=1}^n \sum_{i=1}^m w_{nm} R\left(\theta^{(d)}, h_{ij}^{(d)}\right) \quad (16)$$

where,

$$\mathbf{w} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,m} \end{pmatrix}$$

3 Data

For benchmarking purposes, we used the drag coefficient data obtained from the computational fluid dynamics computations of the TUDelft HPC cluster. The data contains both 8- and 16- dimensional shape parameters (input) and corresponding drag coefficient (output). The 8D and 16D data have 500 and 1000 points respectively. The scatter plot showing the distribution of the data is shown in Figure 1 with the bottom row representing the cross-plot of the drag coefficient and the input variables.

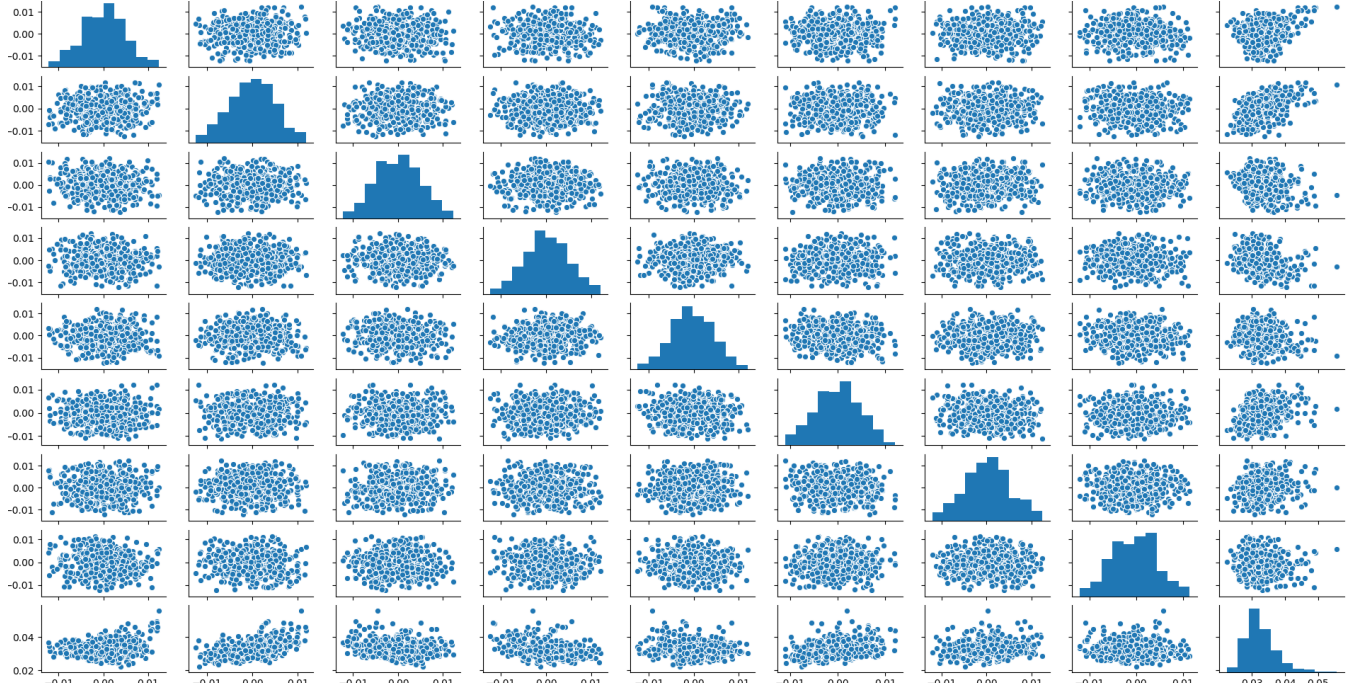


Figure 1: Scatter plot matrix of the 8D data set

4 Approach

In this project report, we used the following approach;

1. Divide the data into training and truth sets (80-20 plan).
2. Draw out samples using Halton sequence ($N_s = 60$) from the training set.
3. Train the model.
4. Predict using truth set inputs.
5. Estimate model performance using different metrics.

For performance measurement, the following metrics are used;

- Normalized root mean square, NRMSE (%)

$$NRMSE = 100 * \sqrt{\frac{1}{m} \sum_i^m \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \quad (17)$$

- Coefficient of determination, R2 score, R^2

$$R^2 = 1 - \frac{\sum_i^m (y_i - \hat{y}_i)^2}{\sum_i^m (y_i - \bar{y})^2} \quad (18)$$

where m is the number of test points (truth set), y_i is the output of the truth set, \hat{y}_i is the predicted result and \bar{y} is the mean of output of the truth set for every i-point.

5 The Halton-based sampling plan

Sampling points used for model training are crucial in the performance of Kriging-based models. It is desirable to have sample data that best describes the given dataset else, the model might have overfitting problems. Because of this, the choice of sampling plan for Kriging / Gaussian processes is not a trivial task. To better improve efficiency of our models, we developed a Halton-based sampling plan to achieve uniformity of data distributions. The algorithm is shown below;

Input: Sampling size, N_s

Output: x_{sample}, y_{sample}

Data: \mathbf{x}, y

Get k from the data

Design a Halton sequence with dimension,k and size, N_s

Normalize \mathbf{x} : x_{norm}

Map the generated sequence on x_{norm} and select the nearest points: $x_{s,norm}$

Denormalize $x_{s,norm}$: x_{sample}

Get the corresponding data from y : y_{sample}

Figure 2 shows the robustness of our developed sampling plan in drawing sampling that best represent the whole dataset.

6 Results

Using the seven models, we bechmarked their performance on the 8D and 16D datasets and also checked the contributions of the single kernels in the multiple kernel methods. Figure 3 and 4 show the contributions of the four individual kernels in the ensemble and CKL methods respectively. From the weighting achieved in the ensemble models, the predominance of the Matérn kernels (3/2 and 5/2) is seen in both 8D and 16D data sets. Similarly, the Gaussian and Matérn 5/2 kernels are more predominant in the CKL formulations. The performance metric of the models are summarized in Figures 5 and 6. The NRMSE metric is more appropriate for the model comparison. It shows that the MCKL model performed best followed by the

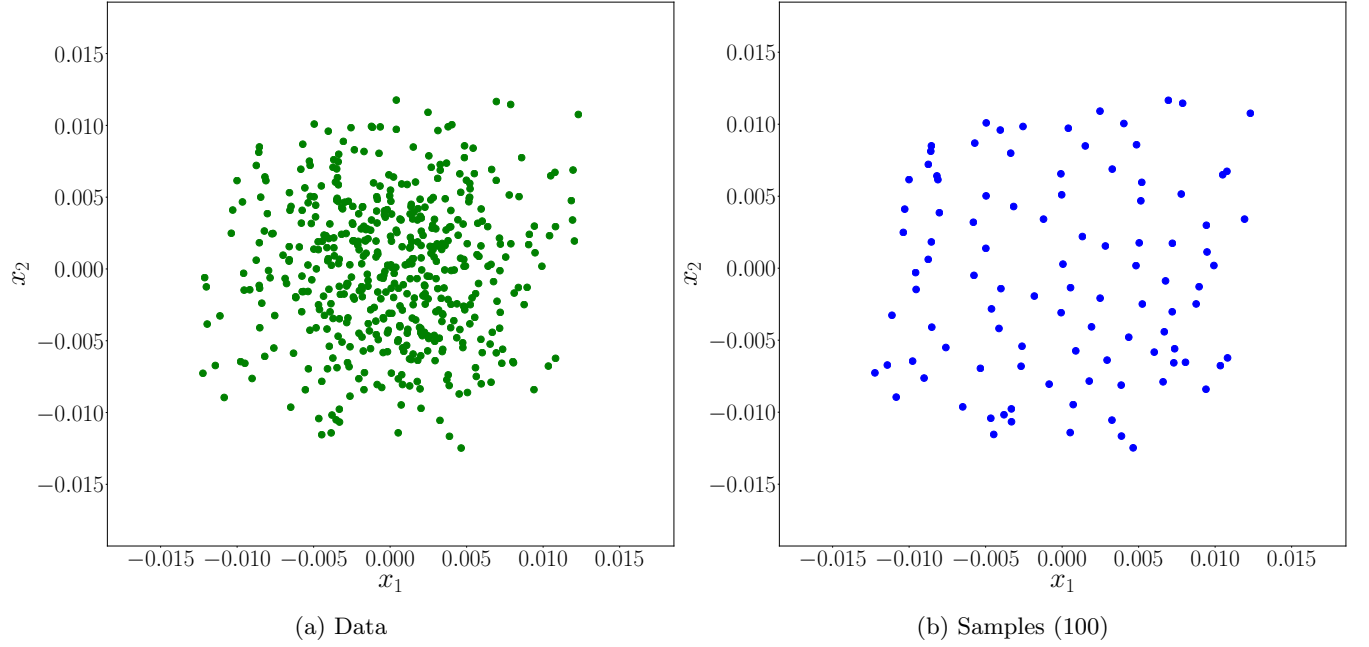


Figure 2: An illustration of the sampling distribution

CKL model in both benchmarkings. However, the training times shown in Figure 7 reveals the high computational time required by the MCKL model.

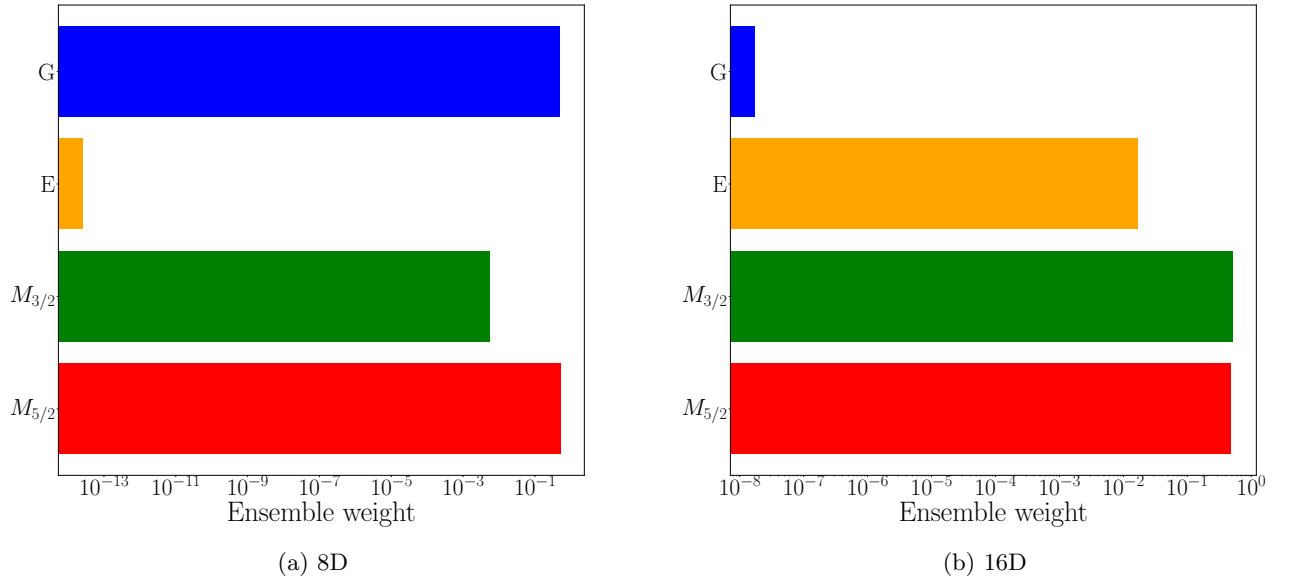


Figure 3: Ensemble weights

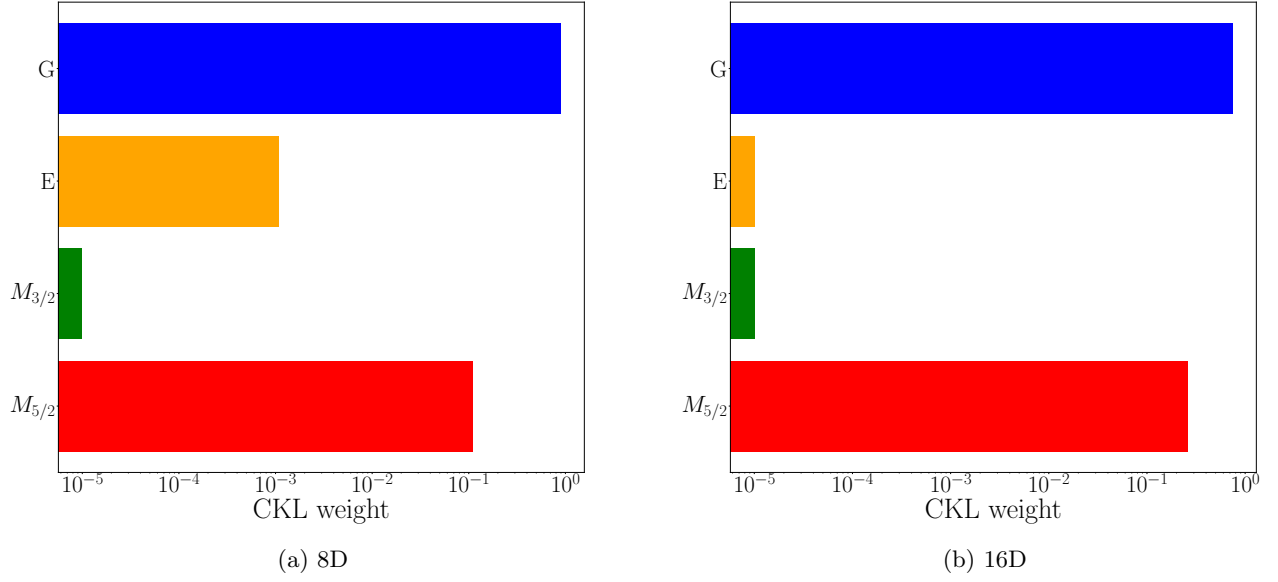


Figure 4: CKL weights

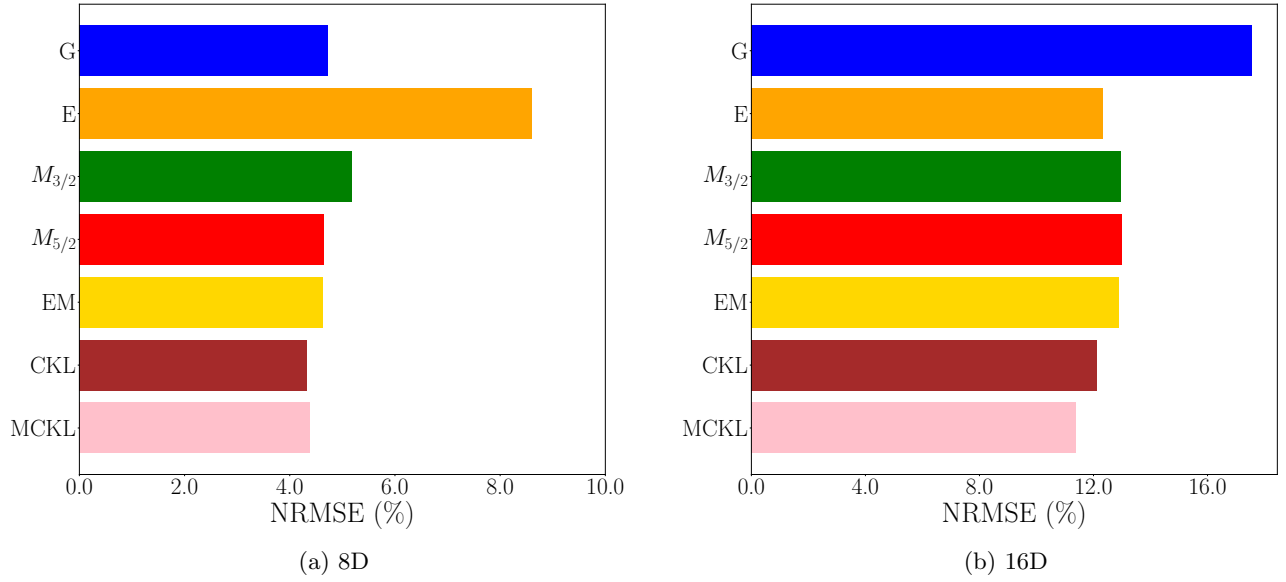


Figure 5: Normalized root mean square errors of predictions

7 Conclusion

7.1 Summary

- Multiple kernel methods are more robust with better performance

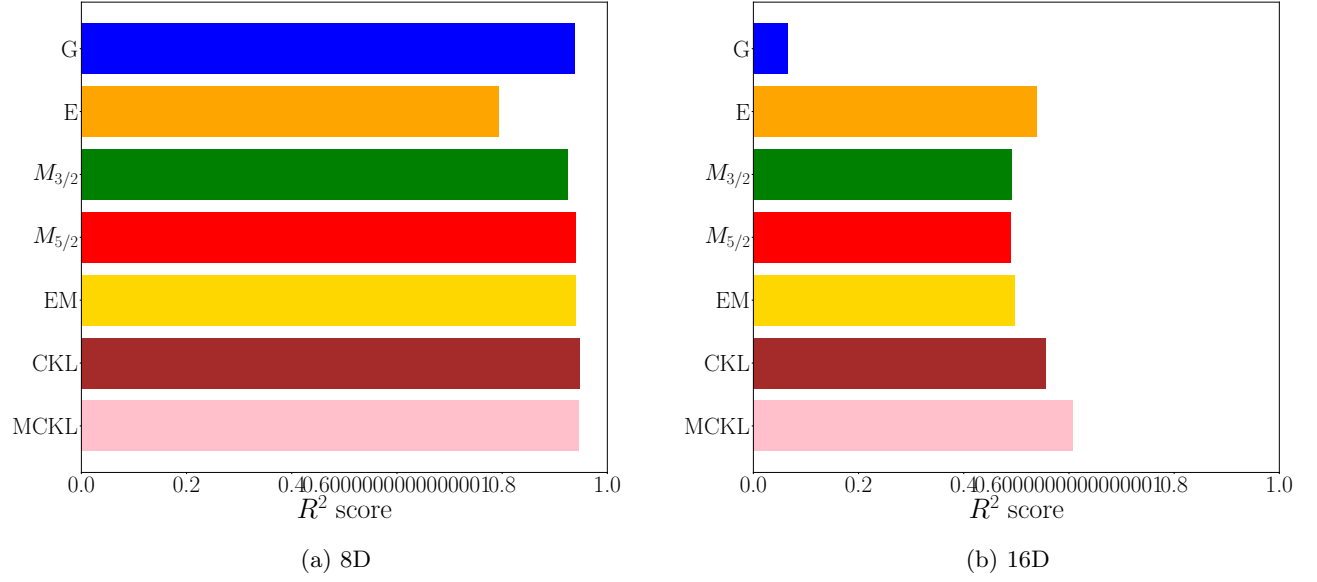


Figure 6: Coefficient of determination score of predictions

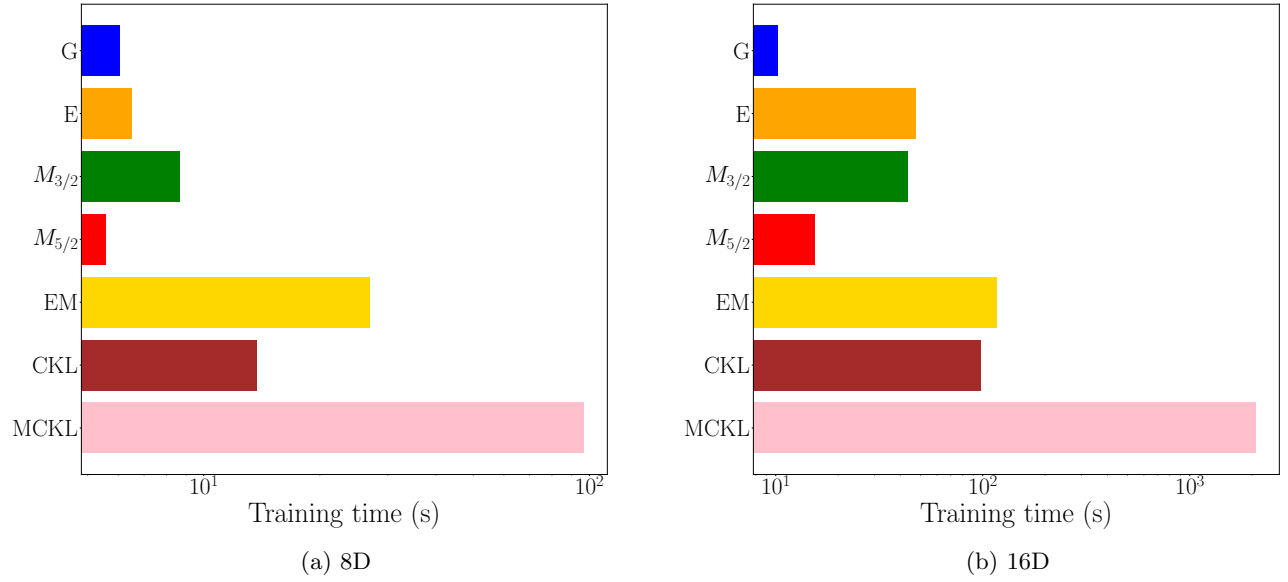


Figure 7: Training time

- MCKL incurred the highest computational cost due to complexity of model
- CKL method is considered best for building models to predict drag coefficients
 - Least training time for multiple kernel methods

- Reasonable model accuracy
- The normalized root mean square method is a more reliable and interpretable metric for model performance comparison.

7.2 Contributions

1. We developed a Halton-based sampling technique.
2. We developed three multiple kernel learning methods based on weighting.
3. We benchmarked performance of methods with single-kernel models.

8 Reproducibility

All the necessary files used to generate this report are available at https://github.com/kennybix/MATH_5470.

References

- [1] Jay D. Martin and Timothy W. Simpson. Use of kriging models to approximate deterministic computer models. 43:853–863, 2005.
- [2] Pramudita S. Palar and Koji Shimoyama. Kriging with composite kernel learning for surrogate modeling in computer experiments. 2019.
- [3] Pramudita Satria Palar and Koji Shimoyama. Ensemble of kriging with multiple kernel functions for engineering design optimization, 2018.
- [4] Pramudita Satria Palar and Koji Shimoyama. Efficient global optimization with ensemble and selection of kernel functions for engineering design. 59:93–116, 2019.