Algorithmic development of Kriging-based methods for complex problems via improved kernel and hyperparameter selection

Ph.D. Thesis Defense

Kehinde Sikirulai Oyetunde[1]

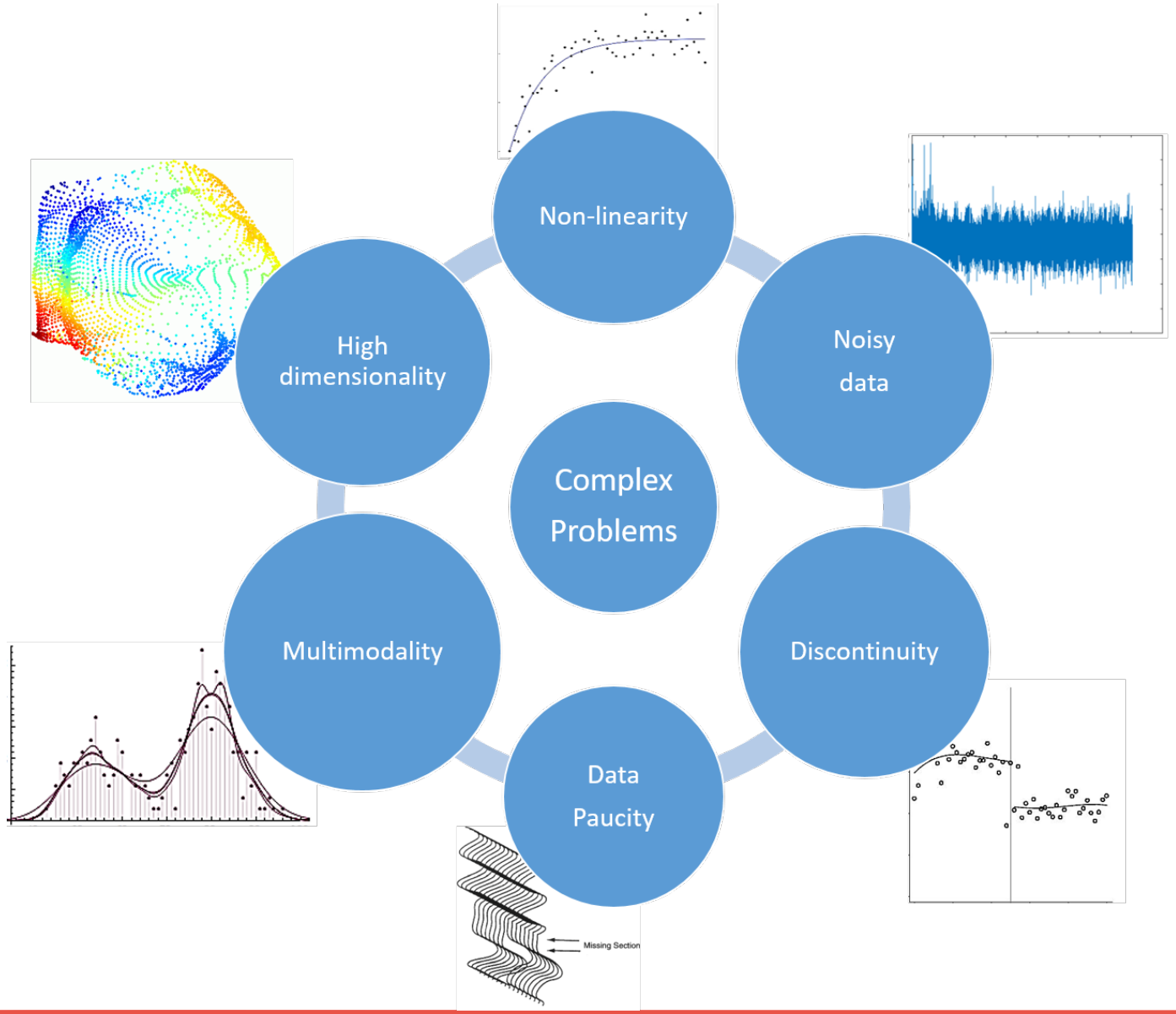
[1]Mechanical and Aerospace Engineering Department,
The Hong Kong University of Science and Technology, Hong Kong

October 28, 2022

Increasing complexity of engineering problems

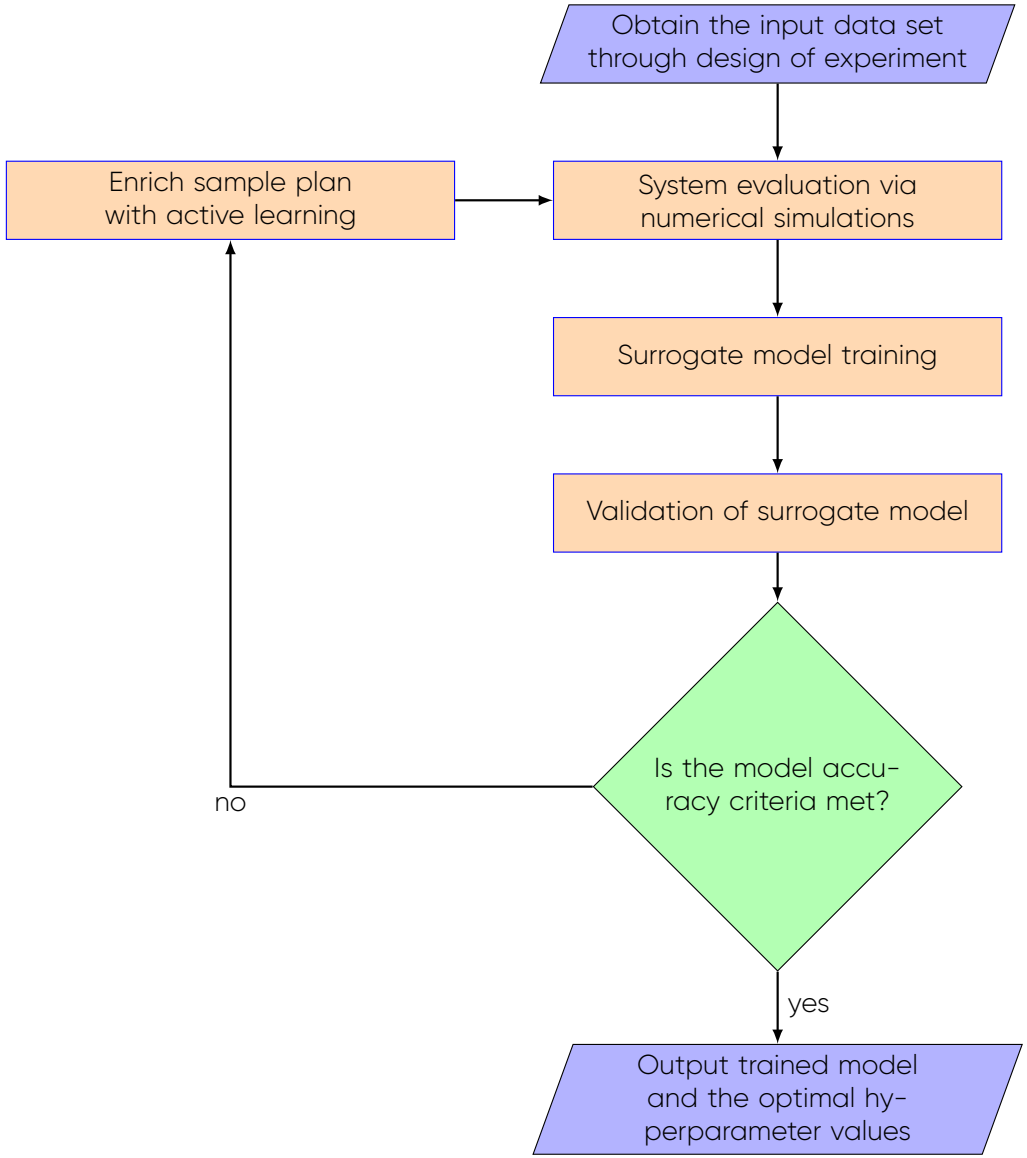
- With ever-growing technological growth, more data are available.
- Sophisticated tools such as FEA and CFD have been developed.
- Increase in the chances of having even more accurate models.
- The improvement in modeling capabilities comes with a cost.
- True function evaluation is costly, because of the inherent intricacies

Characteristics of complex problems



Surrogate-assisted engineering design

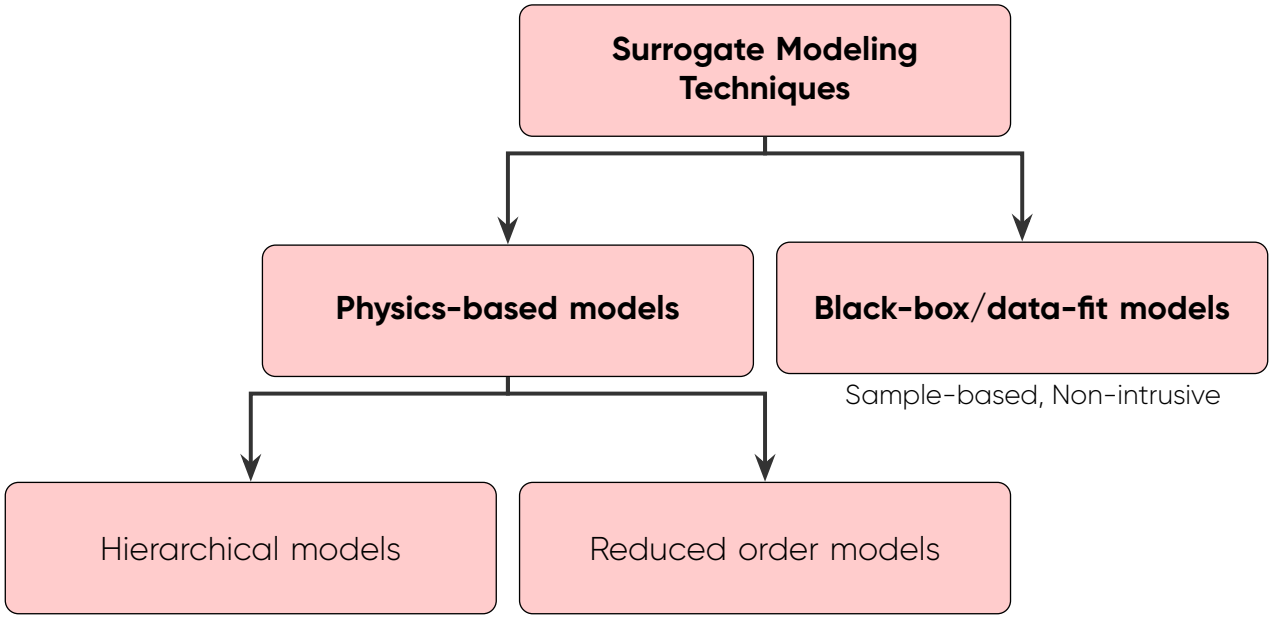
- With surrogates, overall model cost can be greatly reduced (Stork,2020) [a].
- Even more effective solutions can be explored.
- It is important that the surrogates are trusted before deployment.
- When the desired model accuracy is not reached, more points are usually added.



[a] Open Issues in Surrogate-Assisted Optimization

Classification of surrogate models

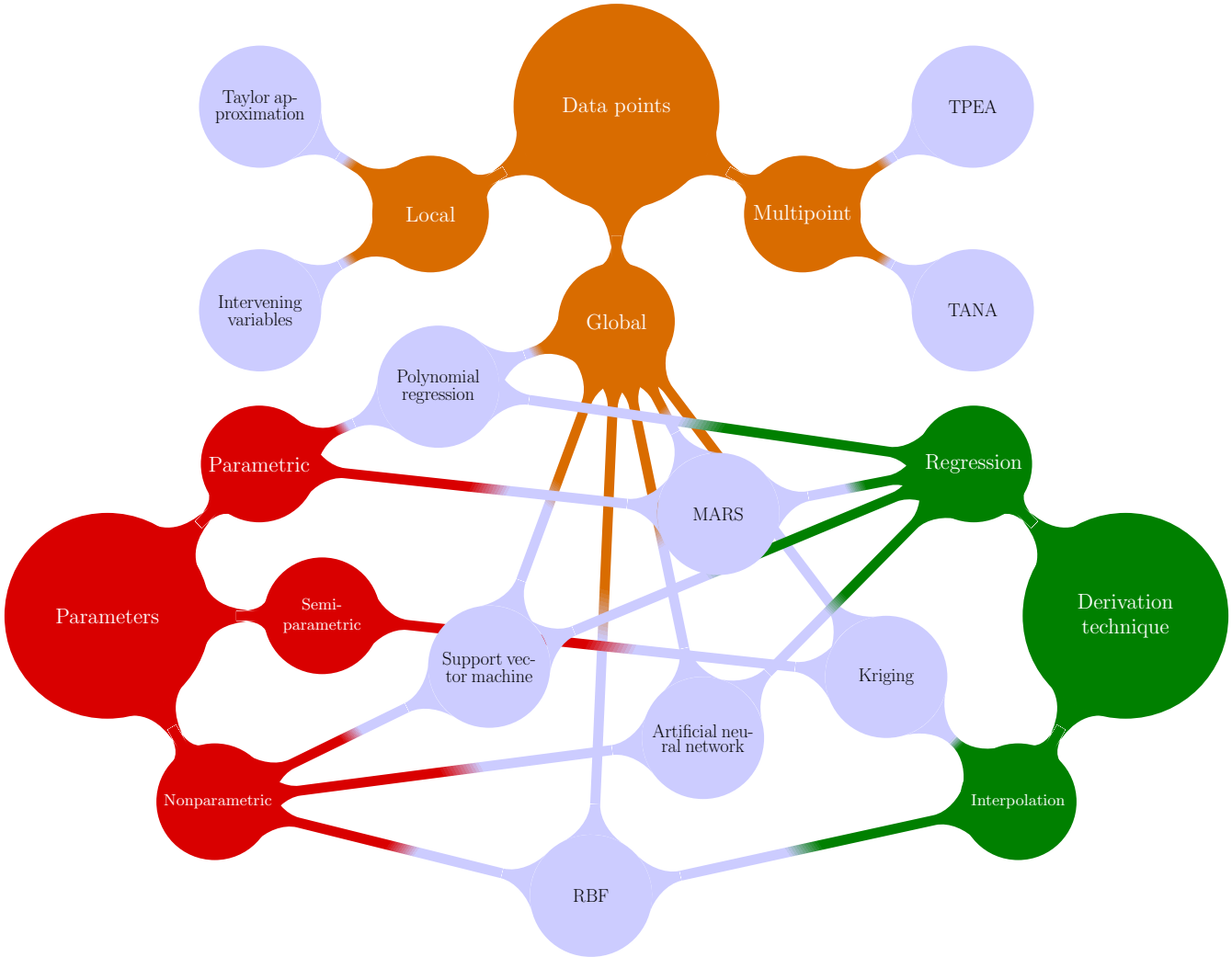
Classification by Eldred *et al.*[1] : data-fit, reduced order, hierarchical models



- **Hierarchical models:** use lower-fidelity models, e.g., higher residual tolerance, coarser grid, simplified physics
- **Reduced order models:** projecting the original model onto a basis that spans a space of lower dimension

[1] M. Eldred and D. Dunlavy. Formulations for surrogate-based optimization with data fit, multifidelity, and reduced-order models. In 11th AIAA/ISSMO multidisciplinary analysis and optimization conference, page 7117, 2006

Black-box surrogate models



R. P. Liem. Multimission Fuel-Burn Minimization in Aircraft Design: A Surrogate-Modeling Approach. University of Toronto

Why Kriging?

- Kriging is one of the commonly used methods, especially in engineering applications (Palar *et al.*, 2018)[2].
- The ease of estimating uncertainty from the output variance makes it useful in Bayesian optimization and uncertainty quantification applications (Palar *et al.*, 2019)[3]
- The incorporation of kernels within kriging formulations makes it a flexible predictor (Rasmussen Williams, 2006)[4].
- The gradient information can be easily incorporated into kriging to improve model accuracy and reduce overall uncertainty in prediction (Lockwood Anitescu, 2012)[5].

---

[2] P. S. Palar and K. Shimoyama. Ensemble of kriging with multiple kernel functions for engineering design optimization. In International Conference on Bioinspired Methods and Their Applications, pages 211–222. Springer, 2018

[3] P. S. Palar and K. Shimoyama. Efficient global optimization with ensemble and selection of kernel functions for engineering design. Structural and Multidisciplinary Optimization, 59(1):93–116, 2019.

[4] C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. 2006

[5] B. A Lockwood and M. Anitescu. Gradient-enhanced universal kriging for uncertainty propagation. Nuclear Science and Engineering, 170(2):168–195, 2012

Kriging

Kriging model assumes the deterministic y($\boldsymbol{x}$) is the realization of a stochastic Y($\boldsymbol{x}$) - Sasena (2002)

$$Y(x) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}),$$

- **Model parts**
  - Global term: $\mu(\boldsymbol{x})$
  - Stochastic term: $Z(\boldsymbol{x}) = \boldsymbol{r}(\boldsymbol{x})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{1}\mu)$

where, $R = f(\theta, \kappa, x)$

Global model

- Simple kriging (SK): Known mean
- Ordinary kriging (OK): Constant but unknown mean
- Universal kriging (UK): General polynomial mean

For engineering applications, especially when the trend is not known, it is sufficient to use **ordinary kriging** (Simpson *et al.*,2001)[6]

[6] Simpson, T.W.*et al.*. Kriging models for global approximation in simulation-based multidisciplinary design optimization. AIAA journal, 39(12):2233–2241, 2001.

Qualities of a good kriging model

It is desirable that kriging models have good;

1. model structure (defined by the covariance function)
2. model parameters
   - variance (σ^2)
   - mean (μ)
   - hyperparameter ($\boldsymbol{\theta}$)
3. computational efficiency
   - right amount of sample points
   - optimal training time
4. predictive ability
   - would usually require extra data
   - model validation metrics such as NRMSE, R^2

Common kernels used in kriging framework

The Matérn[7] kernel has a functional form which can be expressed as;

$$\kappa(h, \theta, v) = \frac{1}{2^{v-1}\Gamma(v)} \left(2\sqrt{v}\frac{|h|}{\theta}\right)^{v} K_v \left(2\sqrt{v}\frac{|h|}{\theta}\right)$$

where $v \geq 1/2$ is the shape parameter, Γ is the Gamma function, and K_v is the modified Bessel function of the second kind.

Kernel Types	ν	Expression
Exponential	$\frac{1}{2}$	$\exp -\frac{1}{2}\left(\frac{h}{\theta}\right)$
Matérn 3/2	$\frac{3}{2}$	$\left(1+\frac{\sqrt{3}\lvert h\rvert}{\theta}\right)\exp\left(-\frac{\sqrt{3}\lvert h\rvert}{\theta}\right)$
Matérn 5/2	$\frac{5}{2}$	$\left(1+\frac{\sqrt{5}\lvert h\rvert}{\theta}+\frac{5h^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}\lvert h\rvert}{\theta}\right)$
Gaussian	∞	$\exp -\frac{1}{2}\left(\frac{h}{\theta}\right)^2$

where h is $|x - x'|$ and θ is the length scale

[7] Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning. 2006

Interpretation of θ

- It may be referred to as the *characteristic length scale* - Gramacy (2020) [8]
- It is a measure of the relevance of the input variable - Forrester *et al.*(2008) [9]

For instance,
In an experiment where the quantity of interest (QoI) is **acceleration** of a car

with the inputs as;

1. colour (x_1)
2. engine location (x_2)
3. engine size (x_3)

with $\kappa \propto h/\theta$, then $\theta_1 > \theta_2 > \theta_3$

[8] Robert B Gramacy. Surrogates: Gaussian process modeling, design, and optimization for the applied sciences. 2020

[9] Alexander Forrester, András Sobester, and Andy Keane. Engineering design via surrogate modelling: a practical guide. John Wiley Sons, 2008.

Hyperparameter estimation

- Cross-validation and maximum likelihood estimation (MLE) are the methods of choice for most researchers (Bachoc, 2013)[10].
- MLE method is preferred due to its computational efficiency (Gano 2006)[11].

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N_s/2}|\boldsymbol{R}(\boldsymbol{\theta})|^{1/2}} \exp\left[-\frac{(\boldsymbol{y}-\mathbf{1}\mu)^T R(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\mathbf{1}\mu)}{2\sigma^2}\right]$$

Obtain μ,

$$\mu(\theta) = (\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1})^{-1}(\mathbf{1}^T\boldsymbol{R}^{-1}\boldsymbol{y})$$

Obtain σ,

$$\hat{\sigma}^2 = \frac{1}{N_s}(\boldsymbol{y}-\mathbf{1}\mu)^T\boldsymbol{R}(\boldsymbol{\theta})^{-1}(\boldsymbol{y}-\mathbf{1}\mu)$$

Obtain concentrated log-likelihood function by taking natural logarithm of $L(\boldsymbol{\theta})$

$$LL(\boldsymbol{\theta}) = -\frac{N_s}{2}\ln(2\pi) - \frac{N_s}{2}\ln\hat{\sigma} - \frac{1}{2}\ln|\boldsymbol{R}|$$

[10] F Bachoc. Cross validation and maximum likelihood estimation of hyperparameters of gaussian processes with model misspecification. Comput Stat Data Anal, 66:55–69, 2013

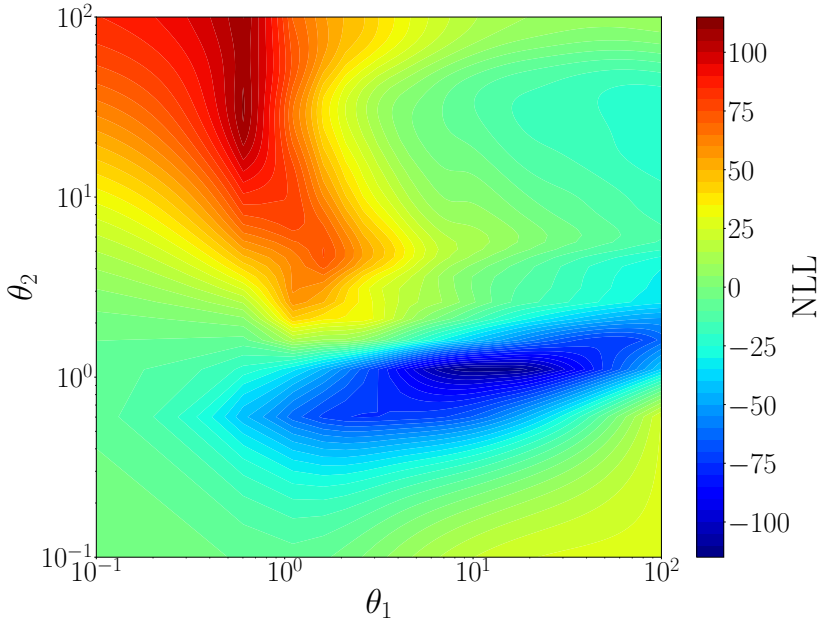
[11] Shawn E Gano, John E Renaud, Jay D Martin, and Timothy W Simpson. Update strategies for kriging models used in variable fidelity optimization. Structural and Multidisciplinary Optimization, 32(4):287–298, jul 2006.

Likelihood estimate and model accuracy

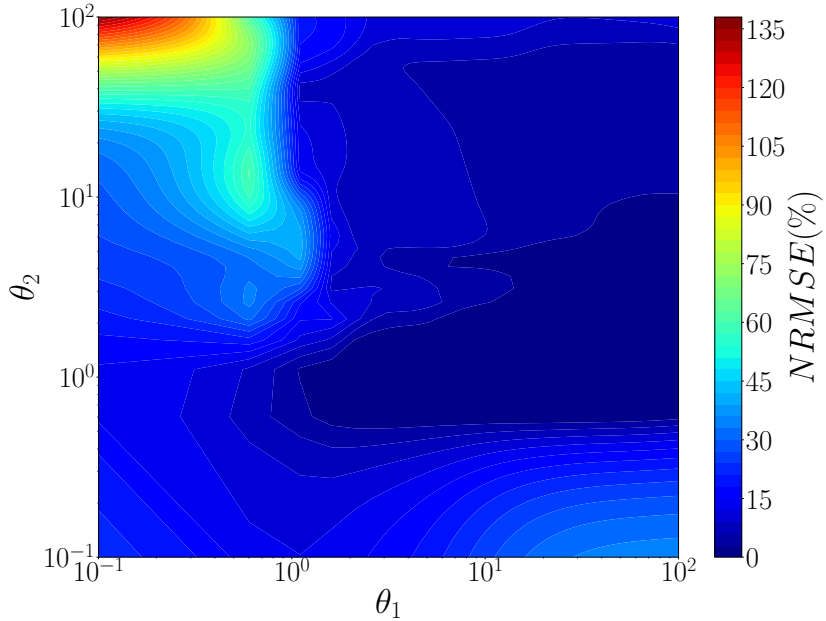
Log-likelihood estimate is positively correlated to the model accuracy.

With the Branin[12] function ($x_1 \in [-5, 10]\,, x_2 \in [0, 15]$),

$$f(x_1, x_2) = \left(x_2 - \frac{5.1x_1^2}{4\pi^2} + \frac{5x_1}{\pi} - 6\right)^2 + 10(1 - \frac{1}{8\pi})\cos(x_1) + 10$$



Negative Log-Likelihood (NLL)



Normalized root mean square error

[12] A. Torn and A. Zilinskas, Global optimization. Springer-Verlag N.Y., Inc., 1989

Issues with kriging

1. Kriging modeling suffers from the *curse of dimensionality*
   - computational complexity of $\mathcal{O}(N_s^3)$ is incurred during model training[13]
   - high dimensional problems may necessitate some dimensionality reduction method[14]
2. Kernel selection
   - usually requires domain expertise
   - poor selections could lead to misspecification[15].
3. Sample selection
   - different problems tend to have different requirements in terms of size and location of samples.
   - with the availability of too many data, there are possibilities of numerical problems[16].

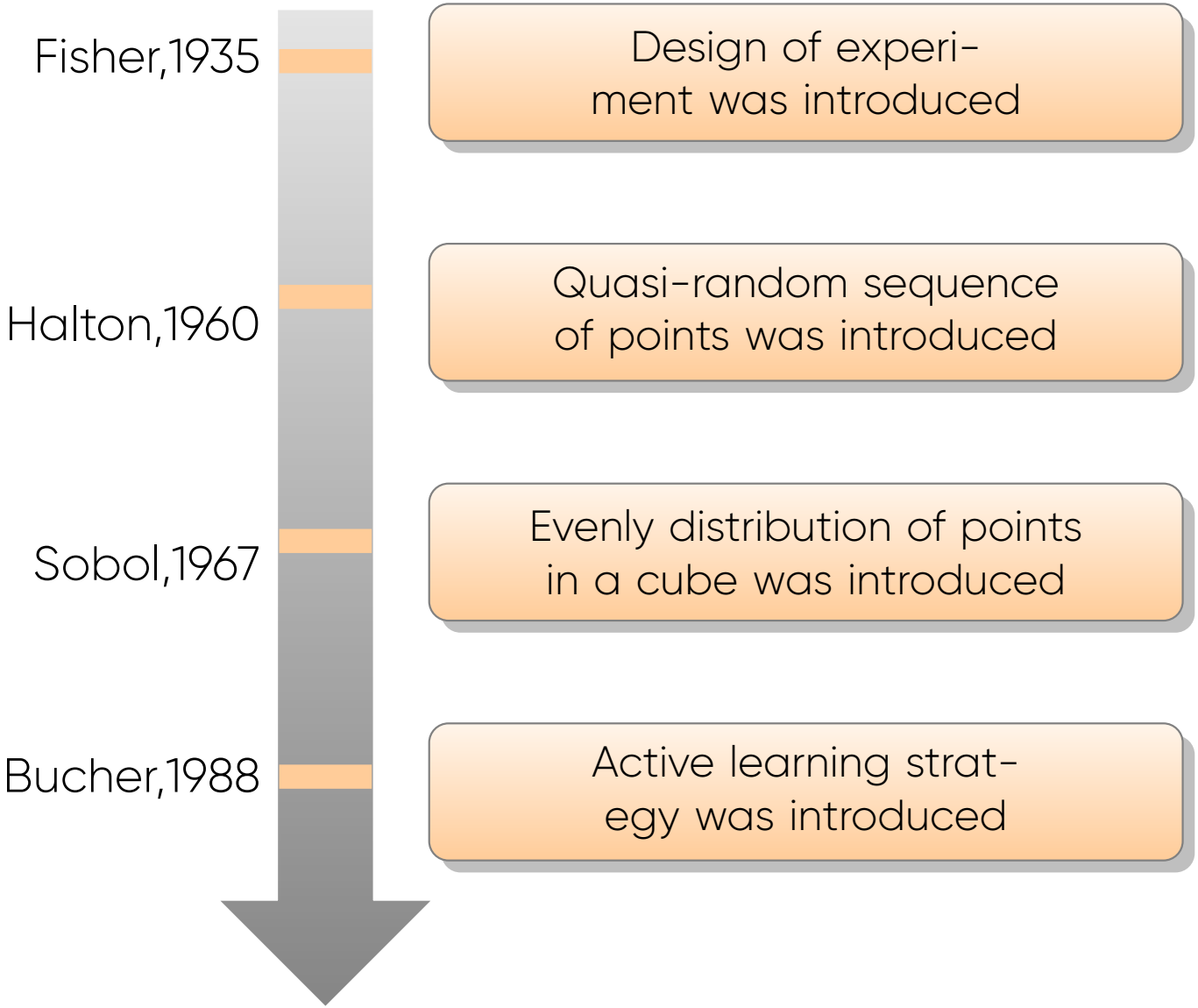
[13] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of global optimization, 34(3):441–466, 2006

[14] Rohit Tripathy, Ilias Bilionis, and Marcial Gonzalez. Gaussian processes with builtin dimensionality reduction: Applications to high-dimensional uncertainty propagation. Journal of Computational Physics, 321:191–223, 2016

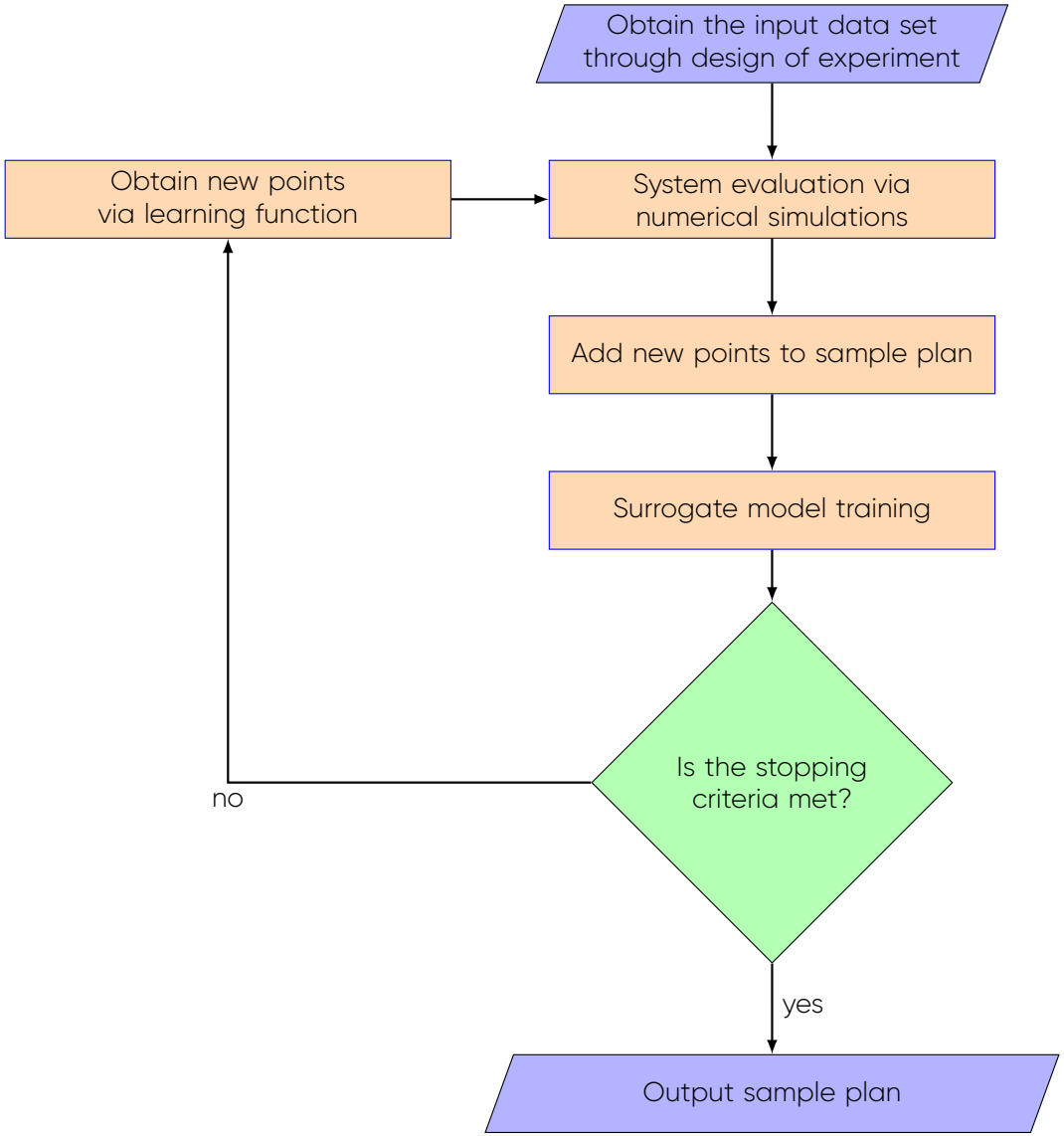
[15] Palar and Shimoyama. Efficient global optimization with ensemble and selection of kernel functions for engineering design. Structural and Multidisciplinary Optimization, 59(1):93–116, 2019

[16] Tinkle Chugh, Alma Rahat, Vanessa Volz, and Martin Zaefferer. Towards better integration of surrogate models and optimizers. In High-Performance SimulationBased Optimization, pages 137–163. Springer, 2020.

History of sample selections



Active Learning strategies



For an effective sample design, a size of $N_s = 10N_d$ is recommended [a]. The learning function can either be based on;

- Exploitation term: $\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2$
- Exploration term: $\hat{s}(\mathbf{x})$

Effective active learning approach should contain both terms and a **tradeoff** (Deschrijver, 2010)

[a] Jones *et al.*, "Efficient Global Functions," Journal of Global Optimization, Vol. 13, No. 4, 1998, pp. 455–492.

Better selections with active learning

Active learning (adaptive sampling) methods aim at improving kriging performance and reducing the points needed to build an accurate model.

1. Eason and Cremaschi (2014)[17] showed that combining adaptive and space-filling techniques could reduce the needed sample points by up to 40% when compared to using purely space-filling methods
2. Liu *et. al.* (2017)[18] proposed the bias-variance decomposition framework to select new points that maximizes the expected prediction error.

- Defining a suitable and robust stopping criterion remains a problem

[17] J Eason and S Cremaschi. Adaptive sequential sampling for surrogate model generation with artificial neural networks. Computers and Chemical Engineering, 68:220– 232, 2014.

[18] H Liu, J Cai, and Y Ong. An adaptive sampling approach for kriging metamodeling by maximizing expected prediction error. Computers and Chemical Engineering, 106:171–182, 2017.

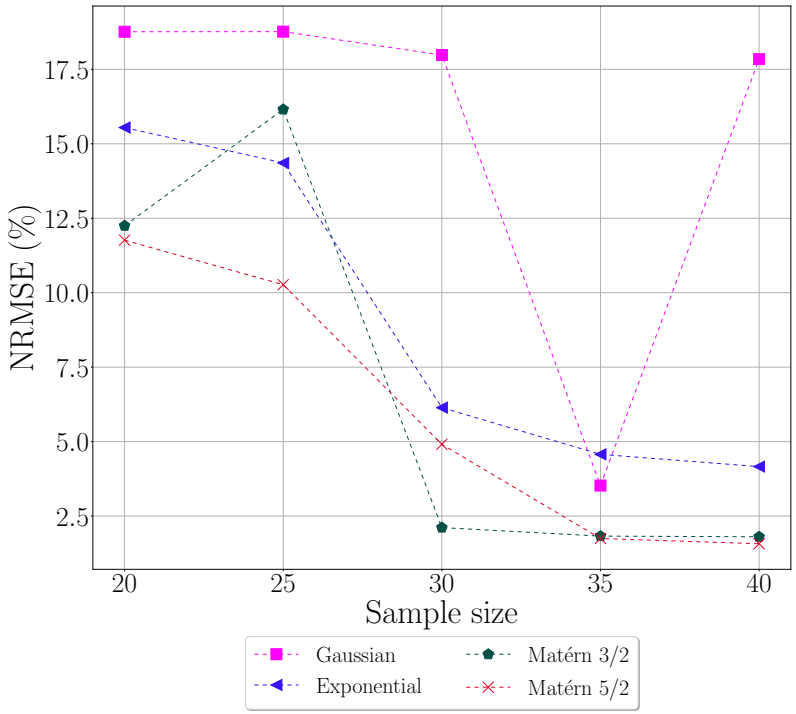
Stopping criteria for active learning can be based on...

- Time constraints
  - Common in industrial applications
  - Could be wasteful
  - No guarantee that the optimum sample points will be selected
- Computational constraints
  - Usually enforced as maximum number of iterations
  - Does not guarantee good performance of resulting sample data
- Accuracy target
  - requires extra data for validation purposes.
  - Extra computational cost incurred by validating the model in each iteration
  - After the solution is met, addition of new point might worsen model performance
- Relative correction between two successive iterations
  - Two successive iterations are usually compared
  - Exhaustive metrics such as cross validation error are used

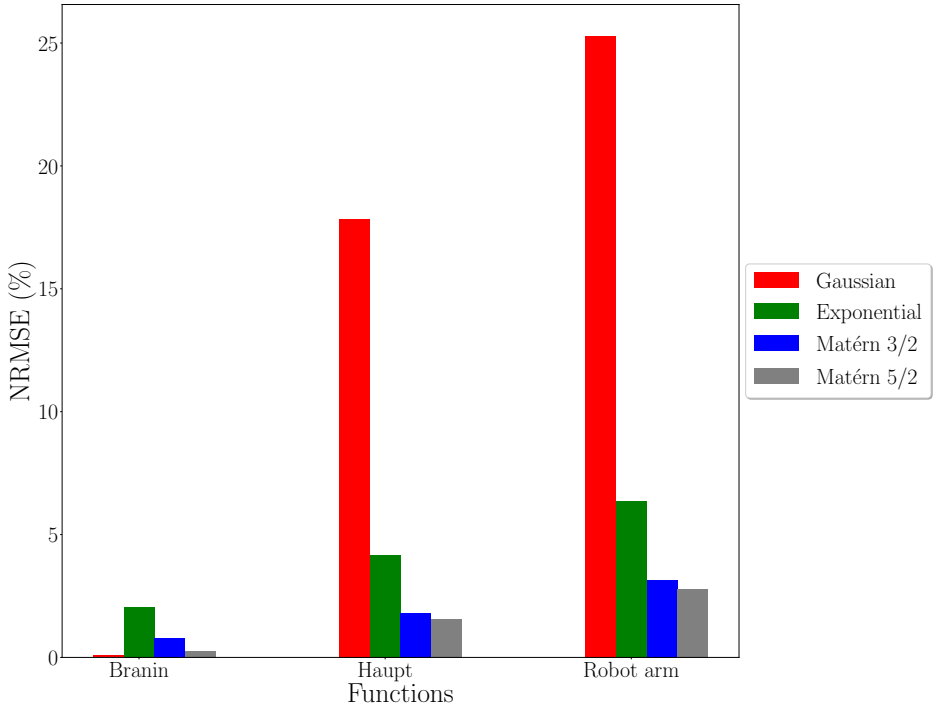
---

Fuhg,J.N. *et al.*. State-of-the-art and comparative review of adaptive sampling methods for kriging. Archives of Computational Methods in Engineering, 28(4):2689–2747, 2021

The choice of suitable kernel varies...



with the sample size (Haupt problem)



with the problem

Hence, it is of great importance to carefully select kernels for our applications

Automatic kernel selections

- Salakhutdinov and Hinton (2007) explored the use of deep learning in the selection of kernels.
  - deep belief net (DBN) to learn covariance kernels for Gaussian processes.
  - used backpropagation through DBN to discriminately fine-tune the covariance kernel, and improved the performance on both regression and classification problems.
- Simpson et al. (2021) proposed a method to identify the kernel for specific applications using transformers.
  - kernel recommendation is performed by a decoder with access to a large vocabulary of primitive kernels.
  - method can predict suitable kernels for a diverse array of real datasets.
- both methods rely on the use of a supercomputer.

Multiple kernel techniques

- Weighted combination of predictions known as ensemble methods[19]
  - Usually a four-stage process
  - Methods are usually time consuming
- Different kernels can also be intrinsically combined efficiently with composite methods[20]
  - Weights are simultaneously trained with the hyperparameters.
  - Negative impact of non-performing kernels is felt more.
  - Possibility of poor performance when the function vary so much across variable dimensions.
- Liem *et. al.* (2019) introduced a brute force approach to select the best kernel combinations across variable dimension.

[19] Pramudita Satria Palar and Koji Shimoyama. Ensemble of kriging with multiple kernel functions for engineering design optimization. In International Conference on Bioinspired Methods and Their Applications, pages 211–222. Springer, 2018.

[20] Pramudita S Palar and Koji Shimoyama. Kriging with composite kernel learning for surrogate modeling in computer experiments. In AIAA Scitech 2019 Forum, page 2209, 2019.

Kriging for high dimensional problems

The inversion of the covariance matrix in the kriging formulation is of a cubic order (Braham *et. al.*, 2014)[21]. This limits the use of kriging to low dimensional problems.

1. Partial least squares coefficient was used in kriging hyperparameter approximations for problems with up to 100 dimensions (Bouhlel *et al.*, 2016)[22].

2. Recently, the use of maximal information coefficient (MIC) in the approximation of kriging hyperparameters was proposed (Zhao *et al.*,2020)[23].

With the methods, huge loss in model accuracy are recorded

---

[21] Braham, Hajer, et al. "Low complexity spatial interpolation for cellular coverage analysis." 2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt). IEEE, 2014.

[22] Bouhlel M.A. *et al.*. Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction. Structural and Multidisciplinary Optimization, pages 935–952, 2016.

[23] Zhao L. *et al.*. An efficient kriging modeling method for high-dimensional design problems based on maximal information coefficient. Structural and Multidisciplinary Optimization, 61(1):39–57, 2020

Research objectives

To improve;

1. scalability of kriging in high-dimensional problems.
   - minimize loss incurred during model reduction.
2. kernel selections in both low and high-dimensional problems.
   - help engineers handle kernel selection effortlessly.
3. sample selection in kriging-based methods.
   - define efficient stopping criteria.

Benchmark functions

Test cases	Characteristics	References
Branin	Smooth, continuous, non-convex	Rajaram *et al.*(2020)
Robot arm	non-linear, symmetric	Hwang & Martins(2018)
Tensor Product		
Hyperbolic Tangent (TPHT)	Multimodal, symmetric	Hwang & Martins(2018)
Cantilever beam	Smooth, continuous, unimodal	Eldred & Burkardt(2009)
Himmelblau	Multimodal, continuous, non-convex	Rajaram *et al.*(2020)
Camel back	Multimodal	Molga and Smutnicki (2005)
Haupt	Multimodal	Haupt *et al.*(2004)
Ackley	Highly multimodal	Adorio & Diliman(2005)

Adorio, E. P., Diliman, U. P. MVF - Multivariate Test Functions Library in C for Unconstrained Global Optimization, 2005.

Molga, M., Smutnicki, C. Test functions for optimization needs,2005.

Eldred, M. S., Burkardt, J. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. In Proceedings of the 47th AIAA Aerospace Sciences Meeting and Exhibit, number AIAA-2009-0976, Orlando, FL (Vol. 123, p. 124).

Haupt, Randy L. Haupt, Sue Ellen. Practical genetic algorithms with CD-Rom (2nd ed.). New York: J. Wiley. ISBN 978-0-471-45565-3, 2004

Rajaram, Dushhyanth, et al. "Empirical assessment of deep gaussian process surrogate models for engineering problems." Journal of Aircraft 58.1: 182-196, 2021.

Hwang, John T., and Joaquim RRA Martins. "A fast-prediction surrogate model for large datasets." Aerospace Science and Technology 75: 74-87, 2018.

High-dimensional benchmark functions

Name	N_d	N_s	Expression
Ellipsoid	20	200	$f(x) = \sum_{i=1}^{20} ix_i^2, x_i \in [-5, 5], i = 1, ..., 20$
Dixon-Price	30	300	$f(x) = (x_1 - 1)^2 + \sum_{i=2}^{30} i(2x_i^2 - x_{i-1})^2, x_i \in [-10, 10], i = 1, ..., 30$
Rosenbrock	40	400	$f(x) = \sum_{i=1}^{39} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right], x_i \in [-5, 10], i = 1, ..., 40$
Griewank[24]	80	500	$f(x) = \sum_{i=1}^{80} \frac{x_i^2}{4000} - \prod_{i=1}^{80} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1, x_i \in [-5, 5], i = 1, ..., 80$

N_s: Number of sample points, N_d: Number of dimensions

[24] Bouhlel, M.A., *et al.*"Efficient global optimization for high dimensional constrained problems by using the kriging models combined with the partial least squares method". Engineering Optimization, 2018

Zhao, Liang, *et al.*"An efficient kriging modeling method for high-dimensional design problems based on maximal information coefficient." Structural and Multidisciplinary Optimization 61.1: 39-57, 2020.

Fu, Chongbo, *et al.*"A distance correlation-based Kriging modeling method for high-dimensional problems." Knowledge-Based Systems 206: 106356, 2020.

Model validation metrics

- Normalized root mean square error (NRMSE):

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

- Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

where,
m = the number of validation points
y_i = Actual value
$\hat{y}_i$ = Predicted value
$\bar{y}$ = Mean of actual values

It is desirable to have low NRMSE $\downarrow$ and high R^2 $\uparrow$

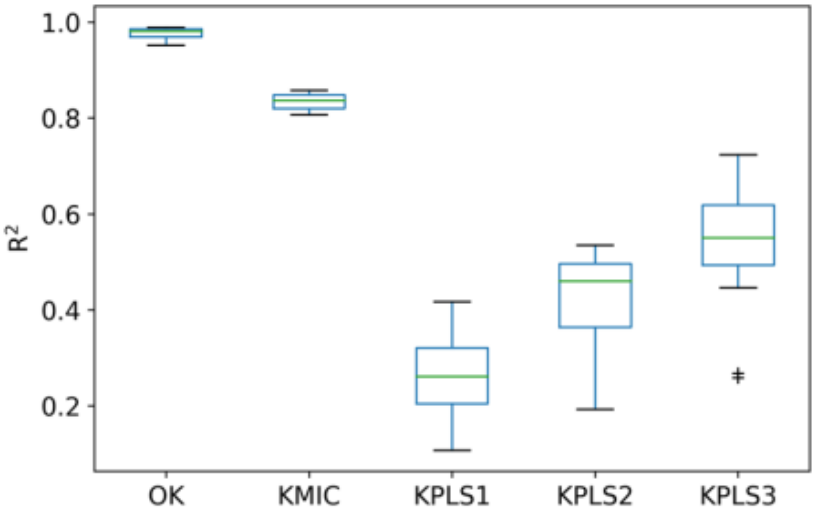
Section 1

Tackling loss of model accuracy in high-dimensional problems

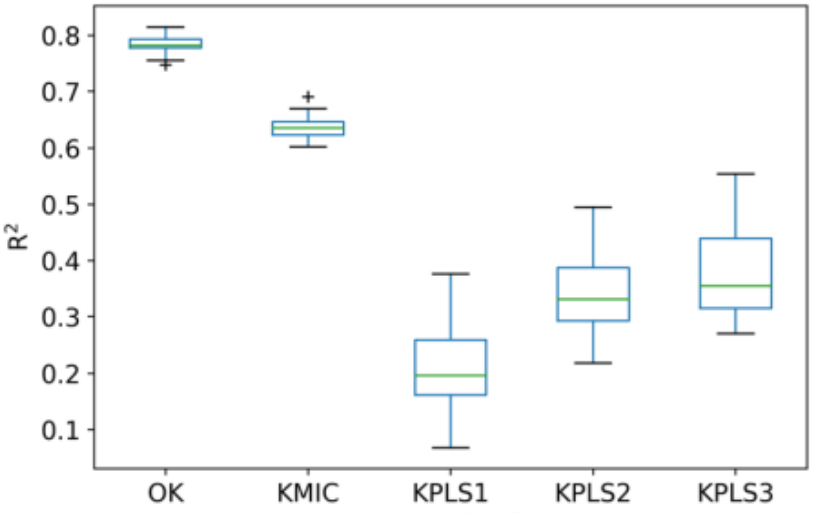
Motivation from literature

With the OK model as the baseline,

- more than 15% loss in model accuracy with the KMIC models in both cases
- poor model accuracy with the KPLS models.



20-D Ellipsoid



30-D Dixon Price

KPLS1: KPLS with **one** principal component

It becomes of importance to **investigate** the loss in model accuracy observed with model reduction.

Zhao, Liang, *et al.*"An efficient kriging modeling method for high-dimensional design problems based on maximal information coefficient." Structural and Multidisciplinary Optimization 61.1: 39-57, 2020.

Approach

1. We propose investigating the model parameter, $\boldsymbol{\theta}$ in both OK and reduced models.
2. It is also of importance to use a function with clear hyperparameter (model parameter) interpretations.
3. Hence, we propose using the Ellipsoid problem[25].
   - With the problem, the next variable is always of **higher significance** than the current variable.

$$x_{n+1} > x_n$$

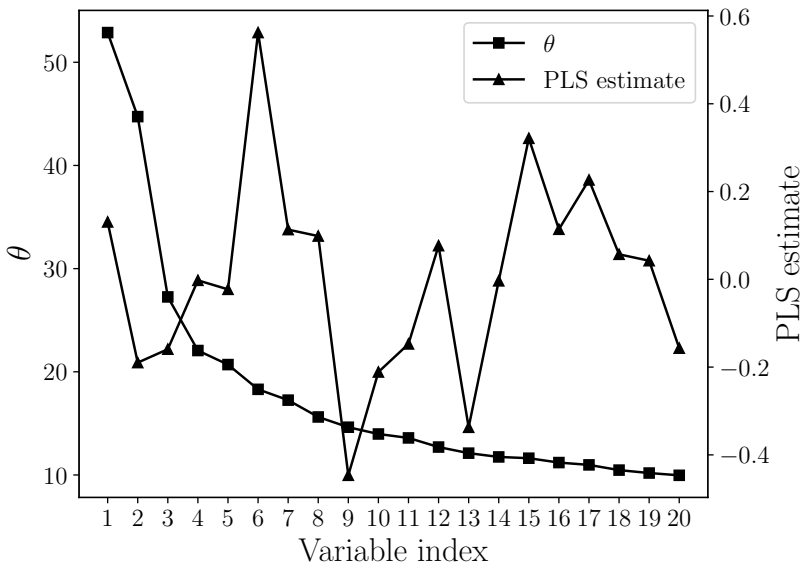
   - In the hyperparameter space, this translates to;

$$\theta_{n+1} < \theta_n$$

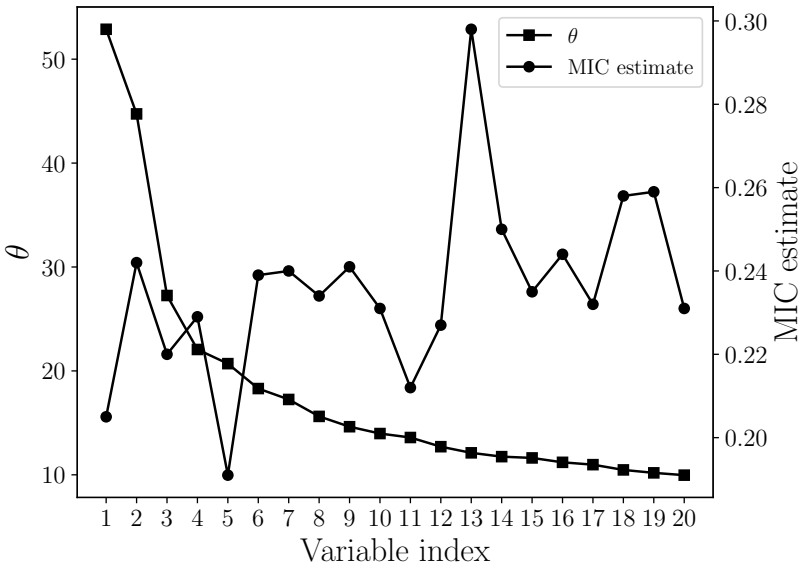
[25] Beyer, Hans-Georg, and Bernhard Sendhoff. "Covariance matrix adaptation revisited–the CMSA evolution strategy–." International Conference on Parallel Problem Solving from Nature. Springer, Berlin, Heidelberg, 2008.

The loss in model structure leads to accuracy loss

- Noticeable changes in the model structure are observed with both PLS and MIC estimates.
- The loss in model accuracy can be attributed to the changes in model structure.



PLS



MIC

Using feature selection metric in hyperparameter approximations

For hyperparameter approximation, it is important that the method captures;

- the relationship of the input variables with the output variable.
- the inter-relationship between input variables.

We propose the use of feature selection methods in achieving this.

- they are used for choosing the most relevant features out of many other features in a given dataset.
- the joint Mutual Information (JMI)[26], minimum redundancy maximum relevance (mRMR)[27] and joint mutual information maximization (JMIM)[28] are common methods.
- with the JMIM method, overestimation of the significance of features can be avoided[29].

---

[26] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. The journal of machine learning research, 13:27–66, 2012.

[27] Radovic M.*et al.*. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics, 18(1):1–14, 2017.

[28] Bennasar M.*et al.*. Feature selection using joint mutual information maximisation. Expert Systems with Applications, 42(22):8520–8532, 2015

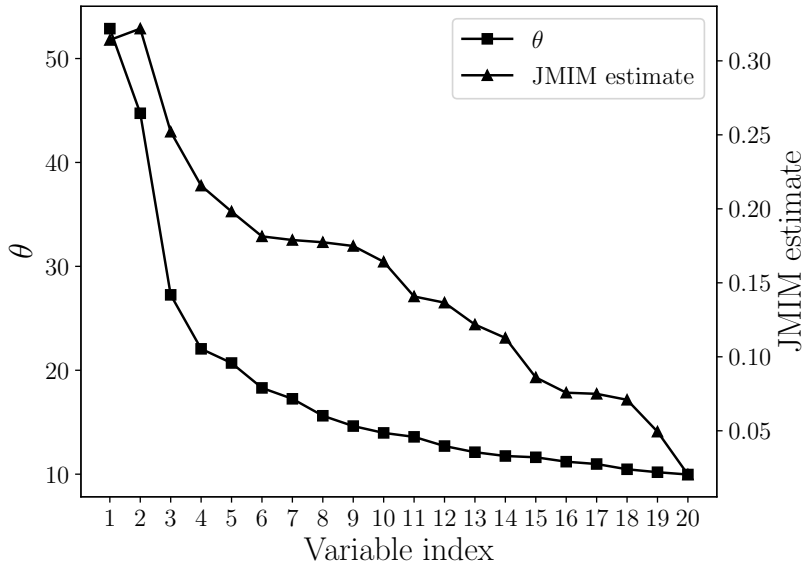
[29] Mao, Y.*et al.*Feature selection based on maximum conditional and joint mutual information. Journal of Computer Applications, 39(3):734, 2019

Kriging with Joint Mutual Information Maximization

1. We obtain the JMIM estimate $\boldsymbol{\lambda}_{JMIM}$
2. We write an approximate equation for $\boldsymbol{\theta}$

   $\boldsymbol{\theta}^f = \omega^f \boldsymbol{\lambda}_{\boldsymbol{JMIM}} + \beta^f \boldsymbol{I}$
3. We then modify the likelihood equations

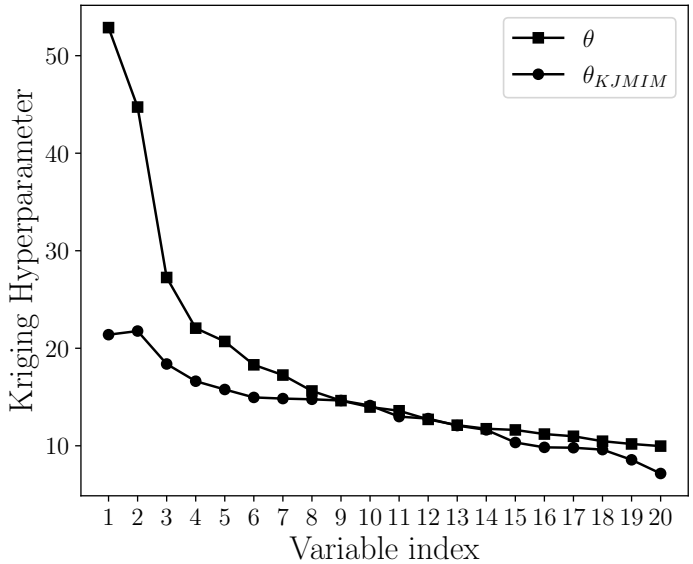
   $\max_{\omega,\beta} \quad -\frac{N_s}{2} ln(2\pi) - \frac{N_s}{2} ln\hat{\sigma}(\omega, \beta) - \frac{1}{2} \ln |R(\omega, \beta)|$



Algorithm Kriging with Joint Mutual Information Maximization (KJMIM)

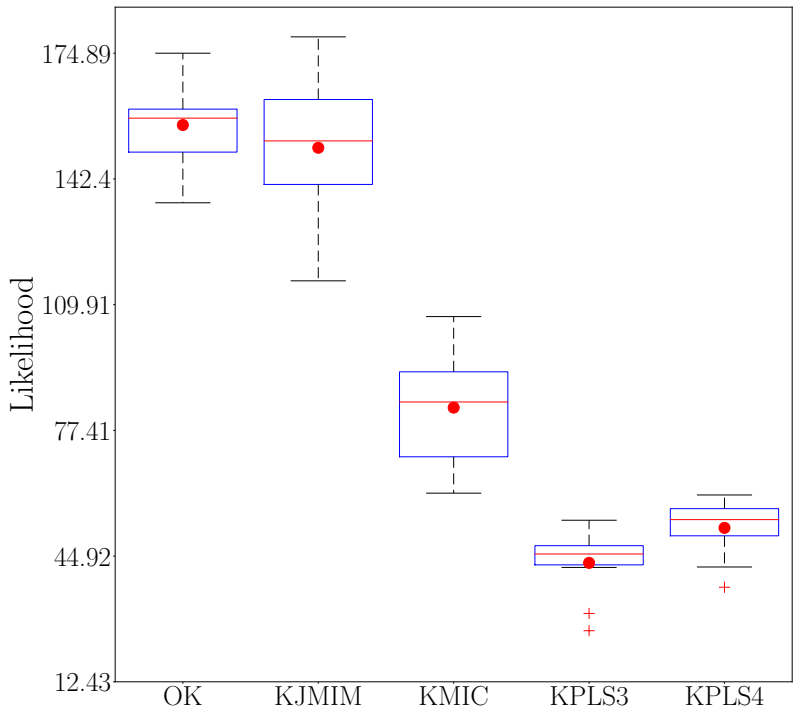
Input: x, y, X
Output: $\theta^f, \omega^f, \beta^f, \hat{\mathbf{y}}$

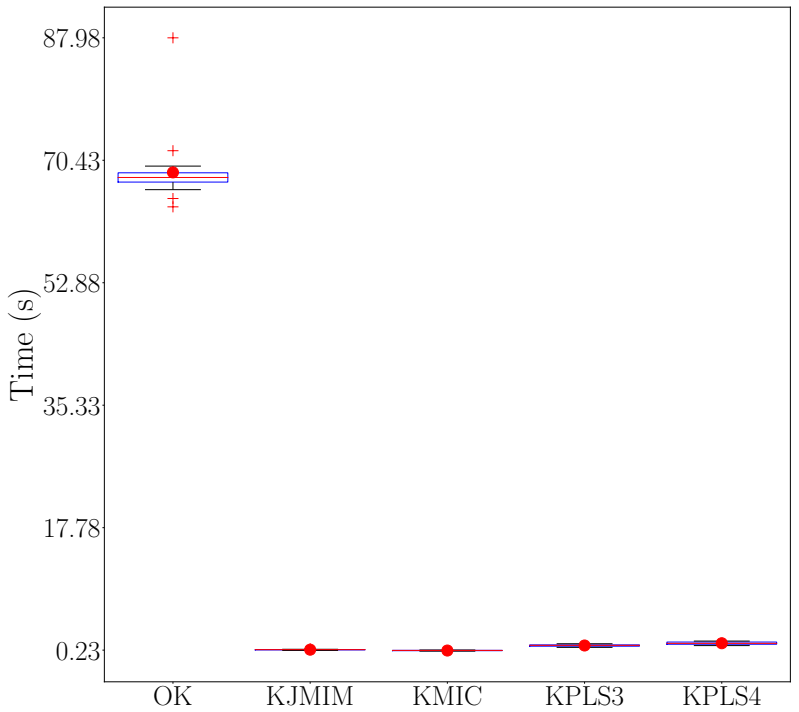
1: Evaluate $\boldsymbol{\lambda}_{\boldsymbol{JMIM}}$
2: Initialize ω, β
3: Assemble $\boldsymbol{R}(\boldsymbol{\omega}, \boldsymbol{\beta}) = \prod_{d=1}^{n} \kappa(\omega, \beta)$
4: Obtain ω^f, β^f by maximizing the likelihood estimate
5: Approximate $\theta^f = \omega^f \boldsymbol{\lambda}_{\boldsymbol{JMIM}} + \beta^f \boldsymbol{I}$
6: Assemble $\mathbf{R}(\theta) = \prod_{d=1}^{n} \kappa\left(\theta_d^f\right)$
7: Compute the kriging weights $\omega = \mathbf{R}^{-1}(\mathbf{y} - \mu)$
8: Estimate $\hat{\mathbf{y}} = \mu + r'(\mathbf{X}) \cdot \omega$



The model structure is conserved with savings in computational time

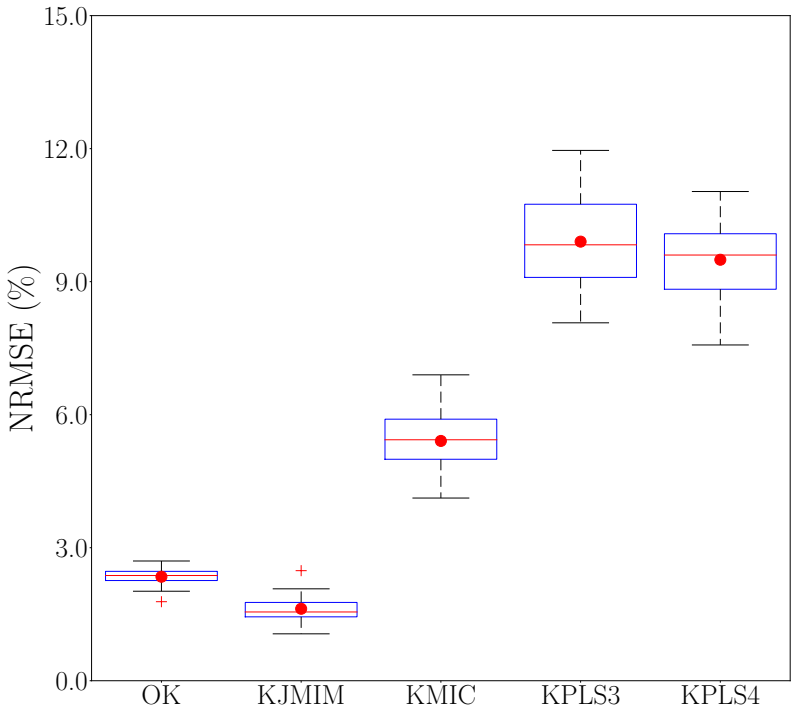
- Clear reduction in training time of the proposed method and existing methods, when compared to the baseline model.
- More than 60% and 50% loss in likelihood estimates of the KPLS and KMIC models respectively.
- Model structure is preserved with the KJMIM model, as shown by the likelihood estimate plot.

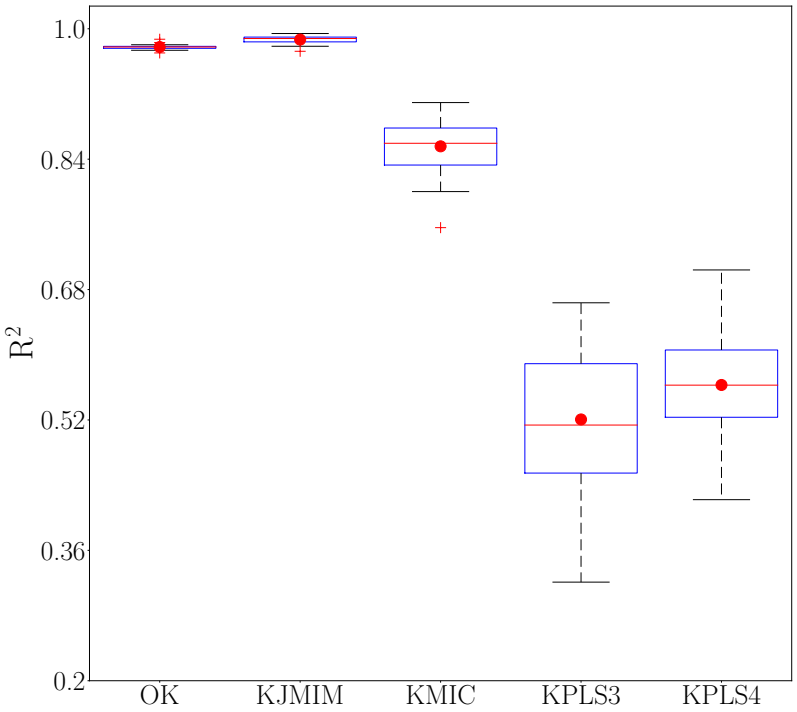




...with no significant loss in model accuracy

- More than 40% loss in the mean R^2 value of the KPLS models, when compared to the baseline model.
- A noticeable improvement in model accuracy is observed with the KJMIM model.





Summary

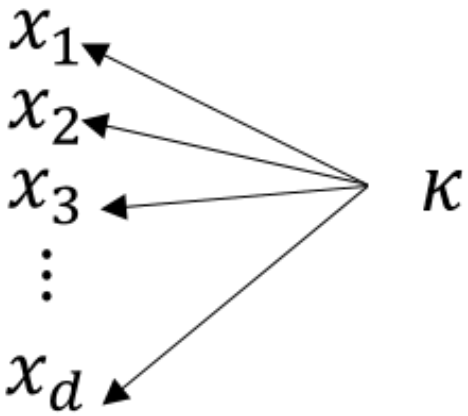
- Feature selection-based techniques can be used for hyperparameter approximations.
- When the model structure is preserved in reduced modeling techniques, the model accuracy of the kriging model can be retained.

Model	R^2	NRMSE (%)	R^2	NRMSE (%)
	Ellipsoid (20-D)		Dixon-Price (30-D)	
OK	0.98	2.35	0.80	6.16
KJMIM	0.99	1.62	0.81	6.30
KMIC	0.86	5.41	0.65	8.48
KPLS-3	0.52	9.90	0.46	10.25
KPLS-4	0.56	9.49	0.52	9.77
	Rosenbrock (40-D)		Griewank (80-D)	
OK	0.78	6.65	0.67	8.55
KJMIM	0.80	6.39	0.80	6.92
KMIC	0.78	6.57	0.77	7.25
KPLS-3	0.72	7.70	0.44	11.90
KPLS-4	0.73	7.52	0.49	11.35

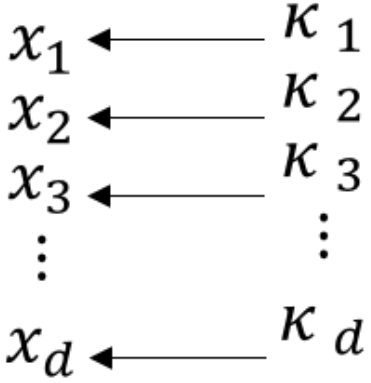
Section 2

Improving kernel selections

Will anisotropic kernel functions be any beneficial?



Single kernel method



Proposed anisotropic method

- How do we obtain the best kernel for each variable dimension?
- How do we make the combinations after discovery?

Kernel-variable combinations

- We propose a new method: Mixed kernel learning method (MIKL).
- A weight matrix is optimized alongside kriging hyperparameters.

Kernels

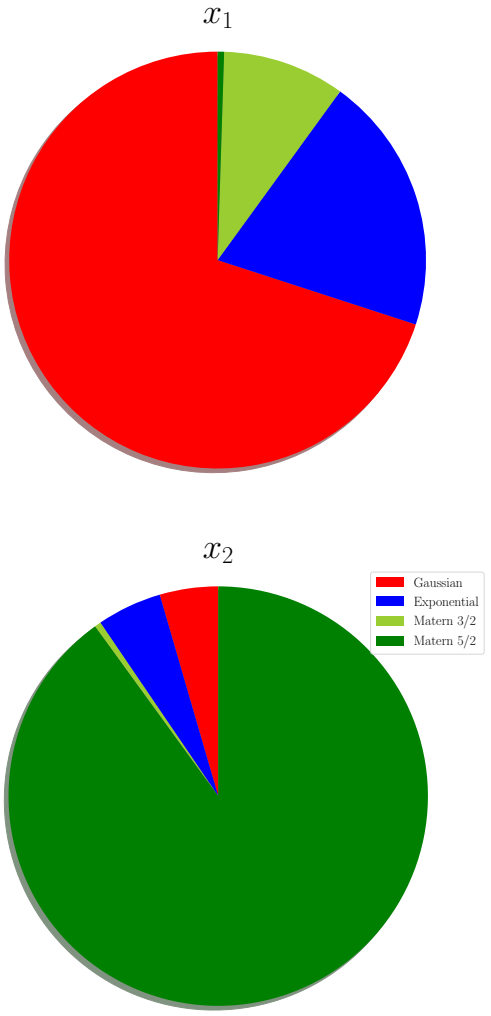
$$\begin{bmatrix} k_1 & k_2 & k_3 & \cdots & k_m \end{bmatrix}$$

$$\text{Variables} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} \Rightarrow \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1m} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2m} \\ w_{31} & w_{32} & w_{33} & \cdots & w_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{d1} & w_{d2} & w_{d3} & \cdots & w_{dm} \end{bmatrix}$$

$$\min_{\mathbf{w},\boldsymbol{\theta}} \quad \frac{N_s}{2}\ln(2\pi) + \frac{N_s}{2}\ln\hat{\sigma}(\mathbf{w},\boldsymbol{\theta}) + \frac{1}{2}\ln|R(\mathbf{w},\boldsymbol{\theta})|$$

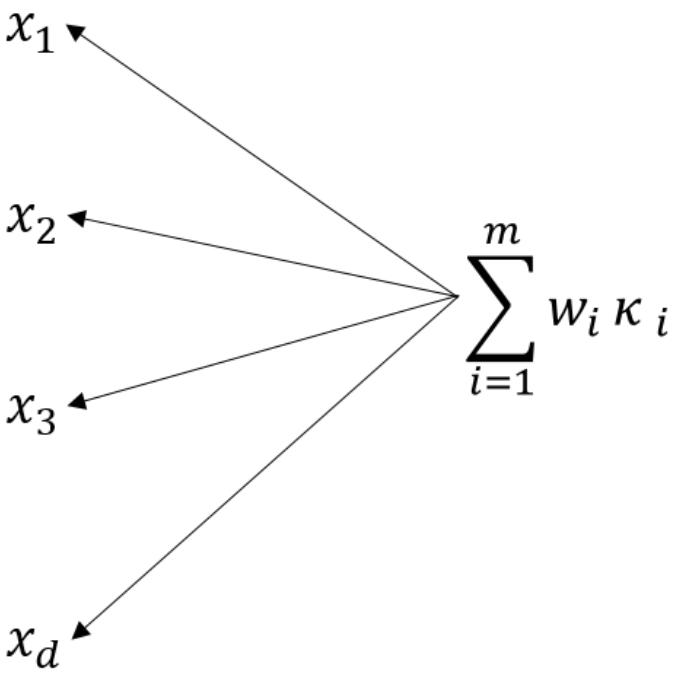
$$\text{s.t.} \quad \sum_{i=1}^{m} w_{di} - 1$$

$$\theta^{(d)} \geq 0$$

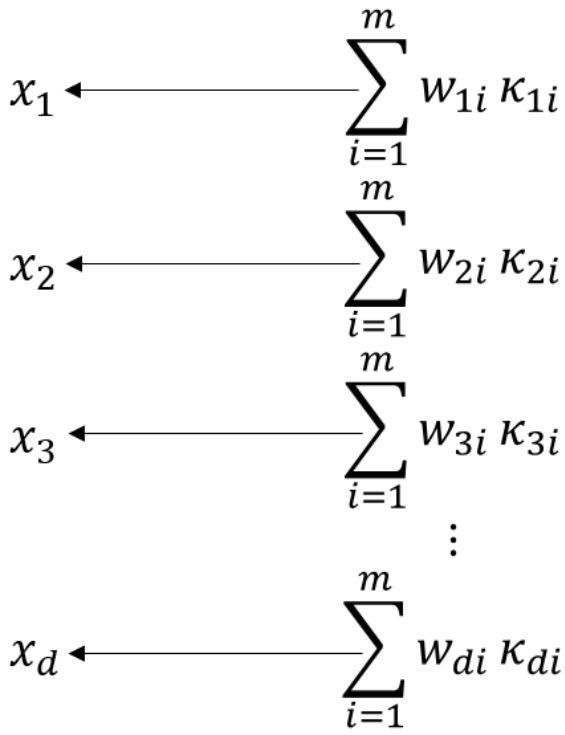


What if we have a unique kernel combination for each variable dimension?

- Multidimensional Composite Kernel Learning (MCKL)



CKL (Palar *et. al., 2019*)[a]



MCKL

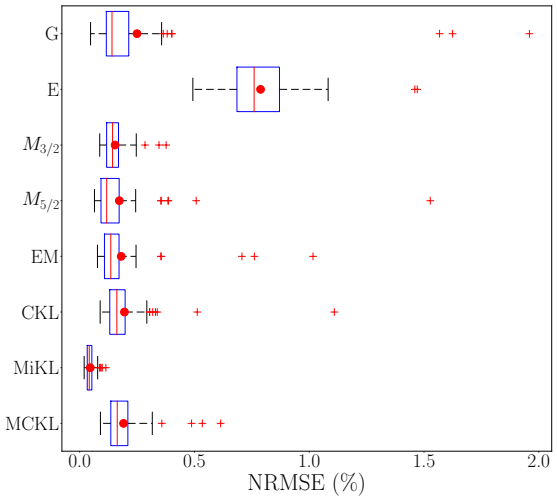
[a] Palar, P.S.*et al.*. Gaussian process surrogate model with composite kernel learning for engineering design. AIAA Journal, 2020.

Benchmark with Mission data

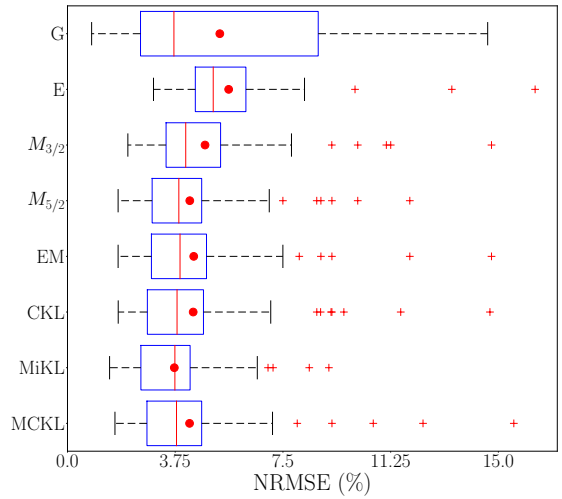
- Four-dimensional Case with CRM Configuration.
- **Output of interest:** C_L,C_D, C_P.

	Lower limit	Upper limit
Mach number	0.0	0.9
Angle of attack (°)	-9	20
Altitude	10	50000
Tail angle (°)	4	40

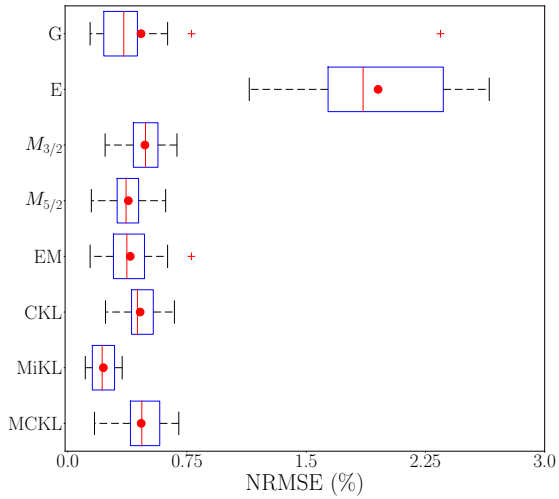
- Multiple kernel methods have good modeling capabilities.
- Noticeable improvement with the MIKL methods especially with the C_l and C_p cases.



Lift Coefficient



Drag Coefficient



Pressure Coefficient

Likelihood profiles and kernel selection

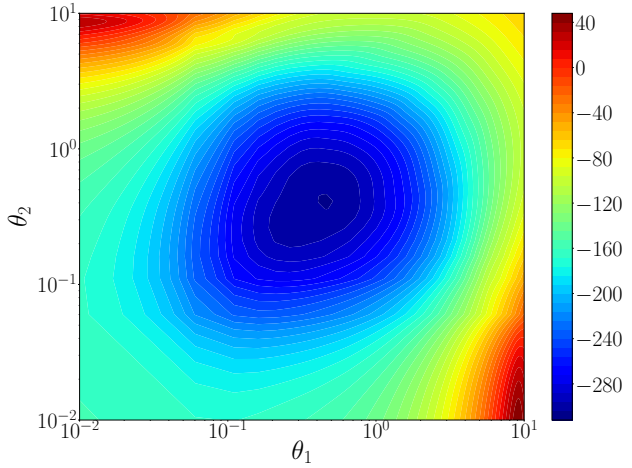
1. Even though multiple kernel models can help automate kernel selection, it is desirable to understand the unique behaviour of each kernel.

2. We propose studying the likelihood profiles since it is of great importance during model training.

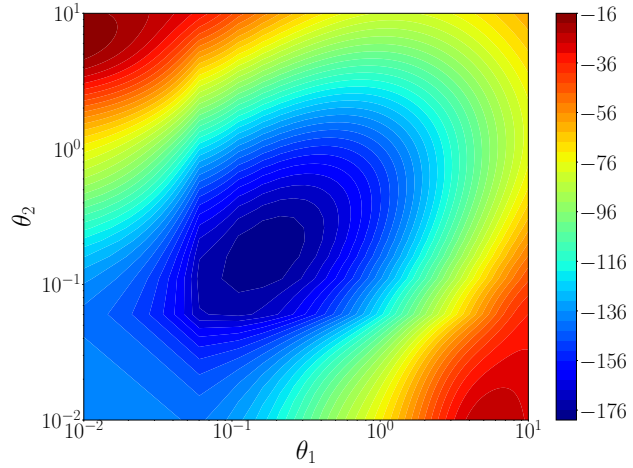
3. Also, it is desirable to propose new methods to effortlessly help with kernel selections.

Negative log-likelihood profiles of cantilever beam problem using different kernels

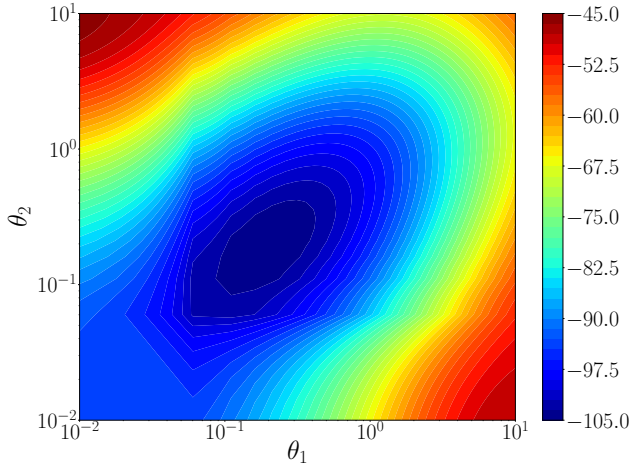
With relatively simple problems, it is easier to obtain global optimal hyperparameter values.



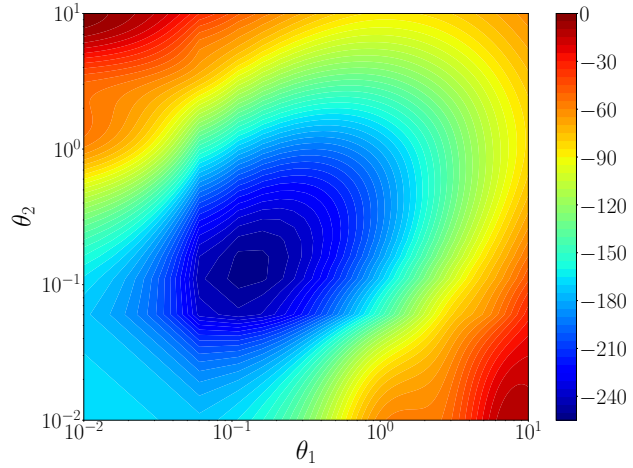
Gaussian



Matérn 3/2



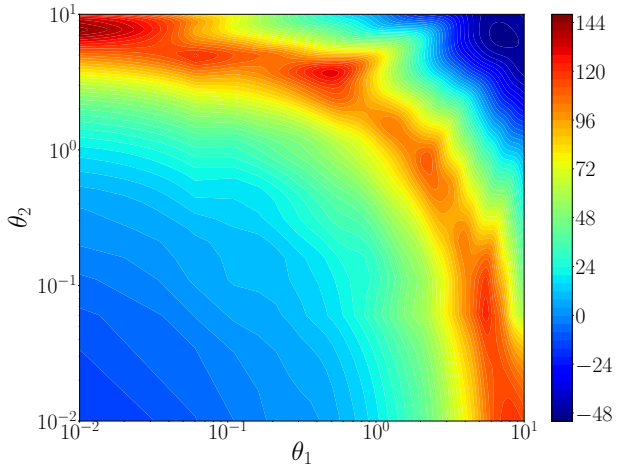
Exponential



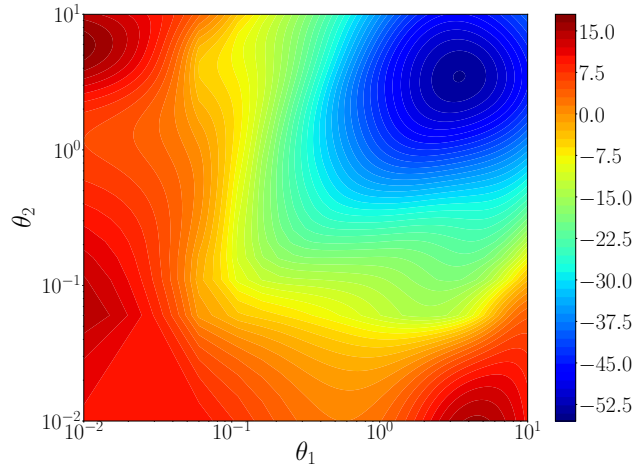
Matérn 5/2

Negative log-likelihood profiles of TPHT problem using different kernels

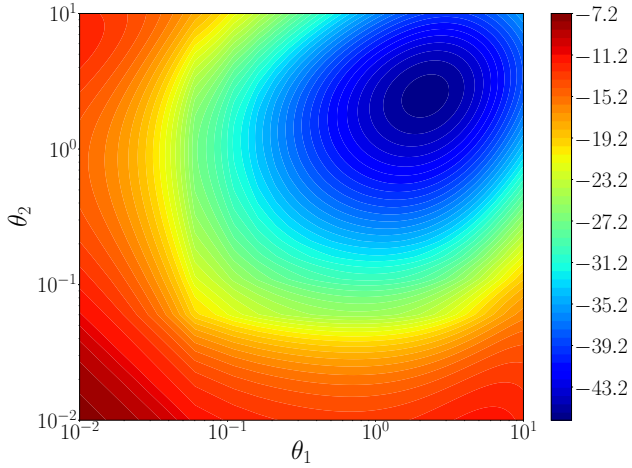
Sub-optimal hyperparameter values are likely especially with the Gaussian and Matérn 5/2 kernels.



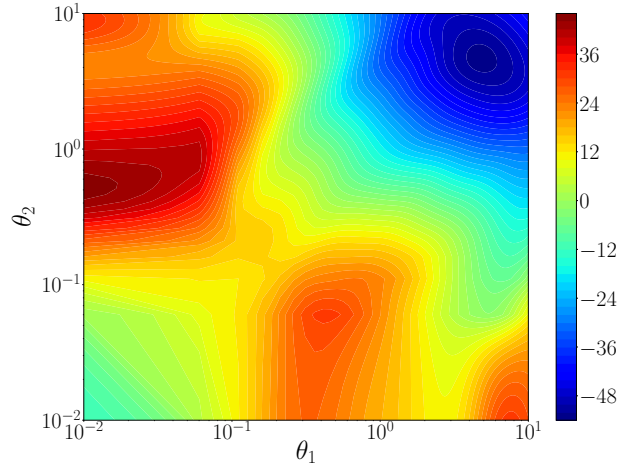
Gaussian



Matérn 3/2



Exponential



Matérn 5/2

Optimal-ν method

- General Matérn kernel is used to obtain the initial hyperparameter values.
- The initial values are used in the second stage.
- Faster convergence and improved model performance can be achieved.

Algorithm Optimal-ν method

Input: x, y, X
Output: θ^f, ν, $\hat{\mathbf{y}}$
1: Initialize θ, ν
2: Assemble $\boldsymbol{R}(\boldsymbol{\theta}) = \prod_{d=1}^{n} \kappa\left(\theta^{(d)}, \nu\right)$
3: Obtain ν^{opt}, θ^f by maximizing the likelihood estimate
4: **if** $\|\nu^f - \nu_{UB}\| \geq \epsilon$ **then**
5: $\quad \nu^f \leftarrow \nu^{opt}$
6: **else**
7: $\quad \nu^f \leftarrow \infty$
8: $\quad$ Obtain θ^f by maximizing the likelihood estimate
9: **end if**
10: Assemble $\mathbf{R}(\theta) = \prod_{d=1}^{n} \kappa\left(\theta_d^f, \nu^f\right)$
11: Compute the kriging weights $\omega = \mathbf{R}^{-1}(\mathbf{y} - \mu)$
12: Estimate $\hat{\mathbf{y}} = \mu + r'(\mathbf{X}) \cdot \omega$

$$\kappa(h, \theta, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\sqrt{\nu}\frac{|h|}{\theta}\right)^{\nu} K_{\nu}\left(2\sqrt{\nu}\frac{|h|}{\theta}\right)$$

$$\max_{\nu,\theta} \quad -\frac{N_s}{2}\ln(2\pi) - \frac{N_s}{2}\ln\hat{\sigma}(\nu,\theta) - \frac{1}{2}\ln|R(\nu,\theta)|$$

$$\text{s.t.} \quad \frac{1}{2} \leqslant \nu \leqslant \frac{6}{2}$$

$$\theta^{(d)} \geqslant 0$$

Optimal-κ method

- Tree-structured parzen in the Optuna framework is used for optimization.
- The kernels are equally optimized in parallel alongside kriging hyperparameters.
- Different kernels and starting points are used during hyperparameter optimization.

Algorithm Optimal-κ method

Input: x, y, **X**, kernel list $\boldsymbol{K} = [\kappa_1, \kappa_2, ...\kappa_p]$, t
Output: θ^f, κ^f, $\hat{\mathbf{y}}$

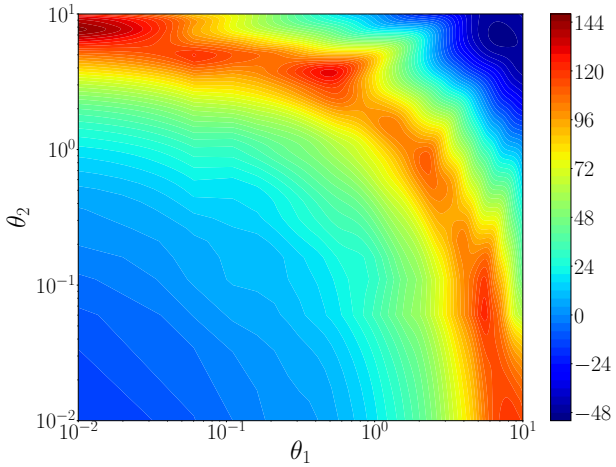
1: Generate t unique combinations
2: $\mathcal{S} = [\,]$; $LL_S = [\,]$
3: **for** each combination **do**
4: Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_t$
5: Assemble $\mathbf{R}(\boldsymbol{\theta}) = \prod_{d=1}^{n} \kappa_t\left(\theta^{(d)}\right)$
6: Obtain $\{\kappa_t^{opt}, \boldsymbol{\theta}_t^{opt}\}$ by maximizing the likelihood estimate
7: Obtain the maximum likelihood estimate, LL from Step 6.
8: $\mathcal{S}$.append($\{\kappa^{opt}, \boldsymbol{\theta}^{opt}\}$)
9: LL_S.append(LL)
10: **end for**
11: $\{\kappa^f, \boldsymbol{\theta}^f\} = \arg\max_{\kappa,\boldsymbol{\theta}} LL_S$
12: Assemble $\mathbf{R}(\boldsymbol{\theta}) = \prod_{d=1}^{n} \kappa^f\left(\theta^{(d)}\right)$
13: Compute the kriging weights $\boldsymbol{\omega} = \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu})$
14: Estimate $\hat{\mathbf{y}} = \boldsymbol{\mu} + r'(\mathbf{X}) \cdot \boldsymbol{\omega}$

$$\max_{\kappa,\theta} \quad -\frac{N_s}{2} ln(2\pi) - \frac{N_s}{2} ln\hat{\sigma}(\kappa,\theta) - \frac{1}{2} ln|R(\kappa,\theta)|$$

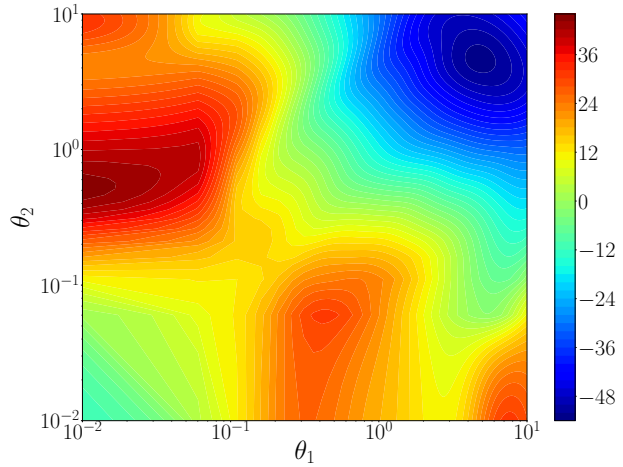
$$\text{s.t.} \quad \theta^{(d)} \geq 0$$

Benchmarking results for algebraic cases NRMSE (%)

- Poorly-performing kernels susceptible to the starting point problem are avoided by the Optimal-κ method.
- Optimal-ν achieves noticeable improvement in the performance of Gaussian kernel models.



TPHT (Gaussian)



TPHT (Matérn 5/2)

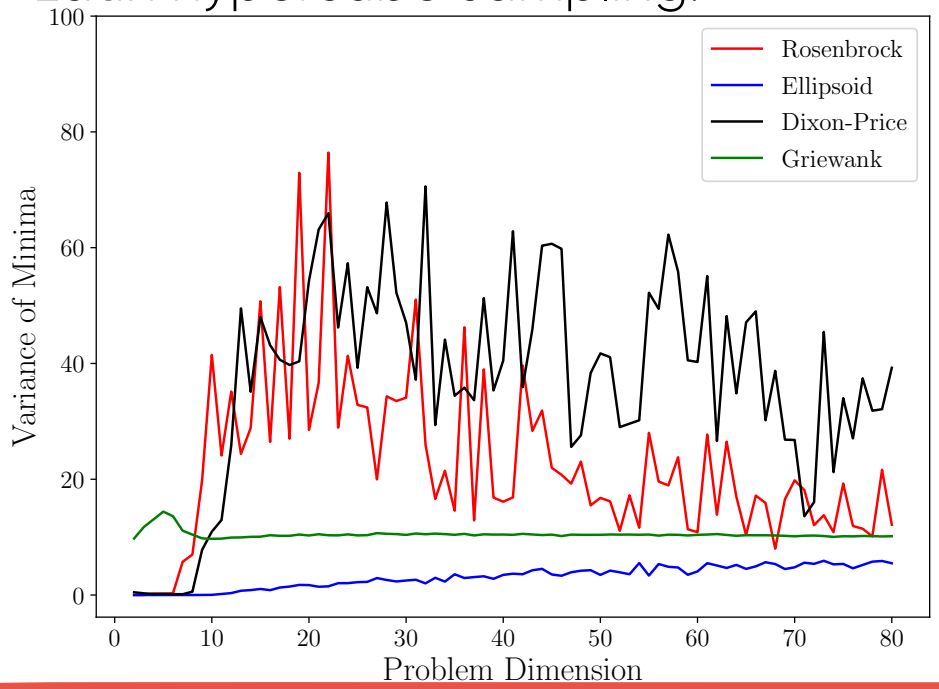
	Branin (2D)	TPHT (2D)	Haupt (2D)
Gaussian	0.0739	12.0957	24.2145
Exponential	9.8358	2.9078	1.9755
Matérn 3/2	2.2200	1.4607	0.5201
Matérn 5/2	0.0469	15.7560	0.3450
Optimal-ν method	0.0651(G)	3.2820(G)	1.2208(G)
Optimal-κ method	0.0662(G)	2.5830(M3)	0.4686(M5)

What's next?

1. With the proposed methods, kernel selection can be automated.
2. We propose looking further into how known problem characteristics can affect kernel selections.
3. Since the TPHT problem used in the previous benchmark is known to be multimodal, we propose looking into the connections.
4. Also, we want to understand the behaviour of reduced models towards kernel selections.

Multimodality and kernel selections

- Variance-based test was designed to capture multimodality of functions.
- In each of the function, we vary the problem dimension from 2 to 80.
- In each problem dimension, 30 initial points are selected with Latin hypercube sampling.



$$\sigma^2_{minima} = \frac{\sum_{i=1}^{p} \left(f_{min} - \bar{f}_{min}\right)^2}{p-1}$$

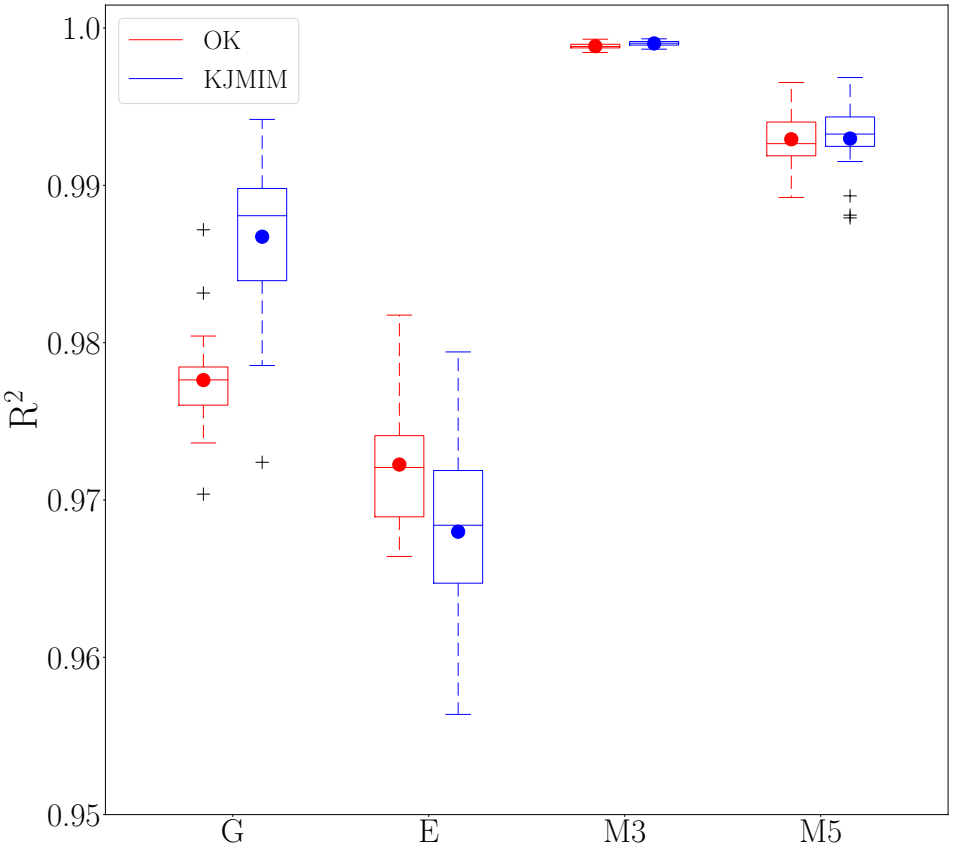
where,

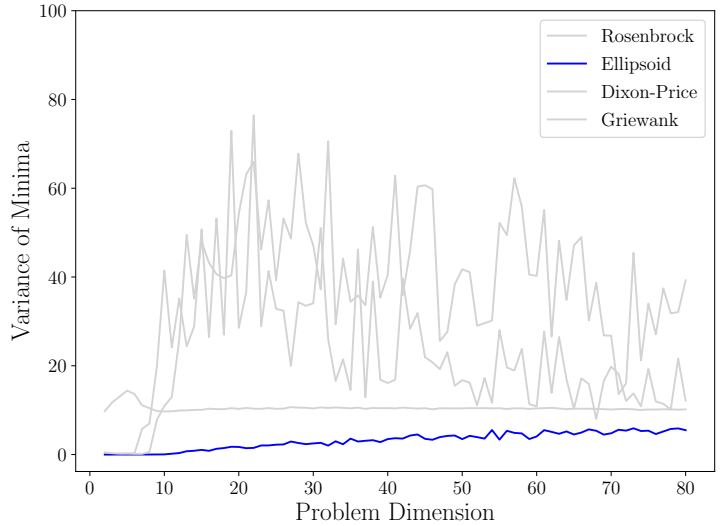
$$\bar{f}_{min} = \frac{\sum_{i=1}^{p} f_{min}}{p}$$

Certain functions change from being unimodal to multimodal

With 20-D Ellipsoid problem, kernel selection is less important

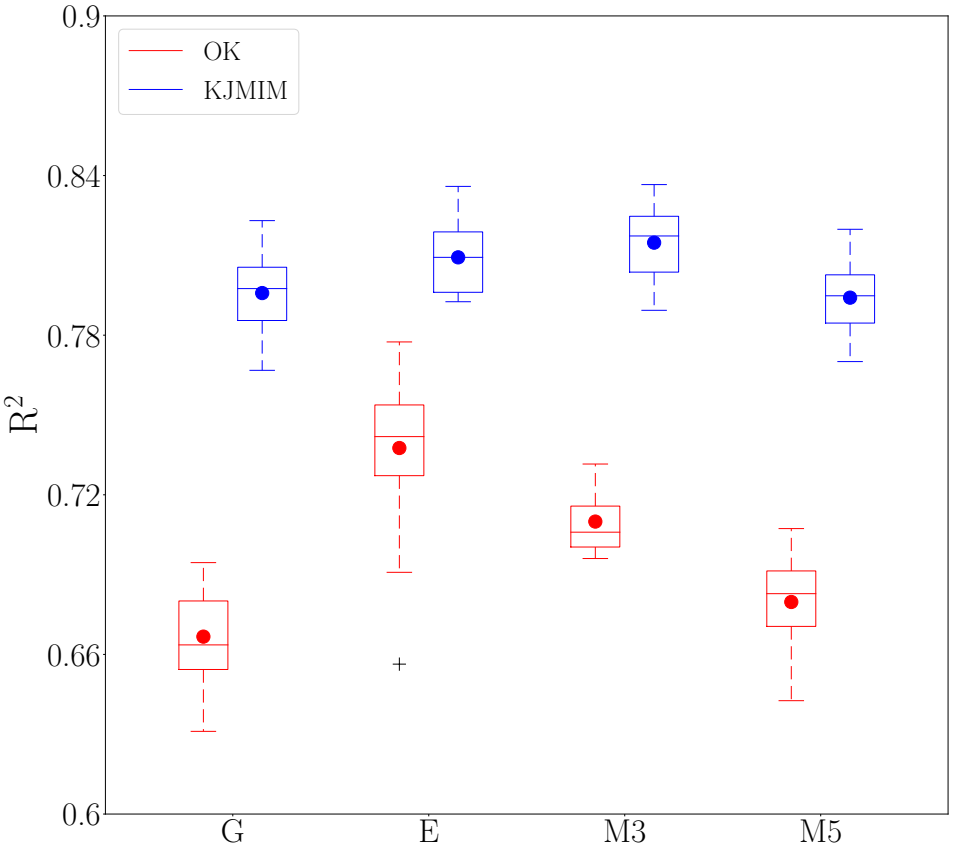
- The 20-D Ellipsoid problem is unimodal.
- All kernels have good modelling ability with both ordinary kriging (OK) and reduced models.

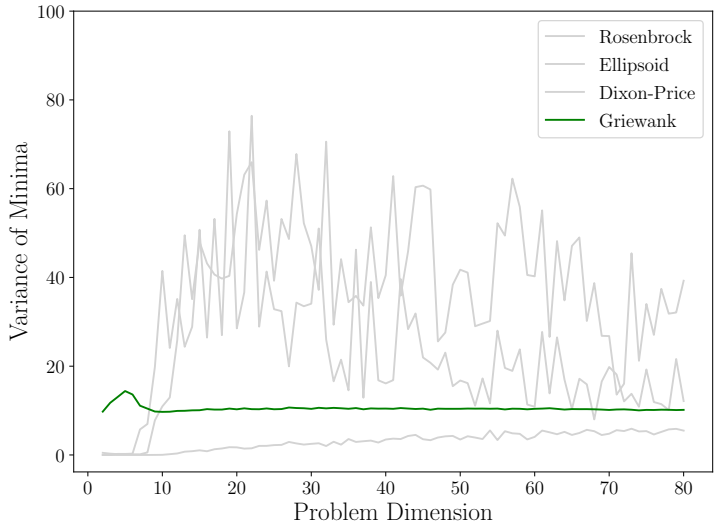




It is important to carefully select the kernel for the 80-D Griewank problem

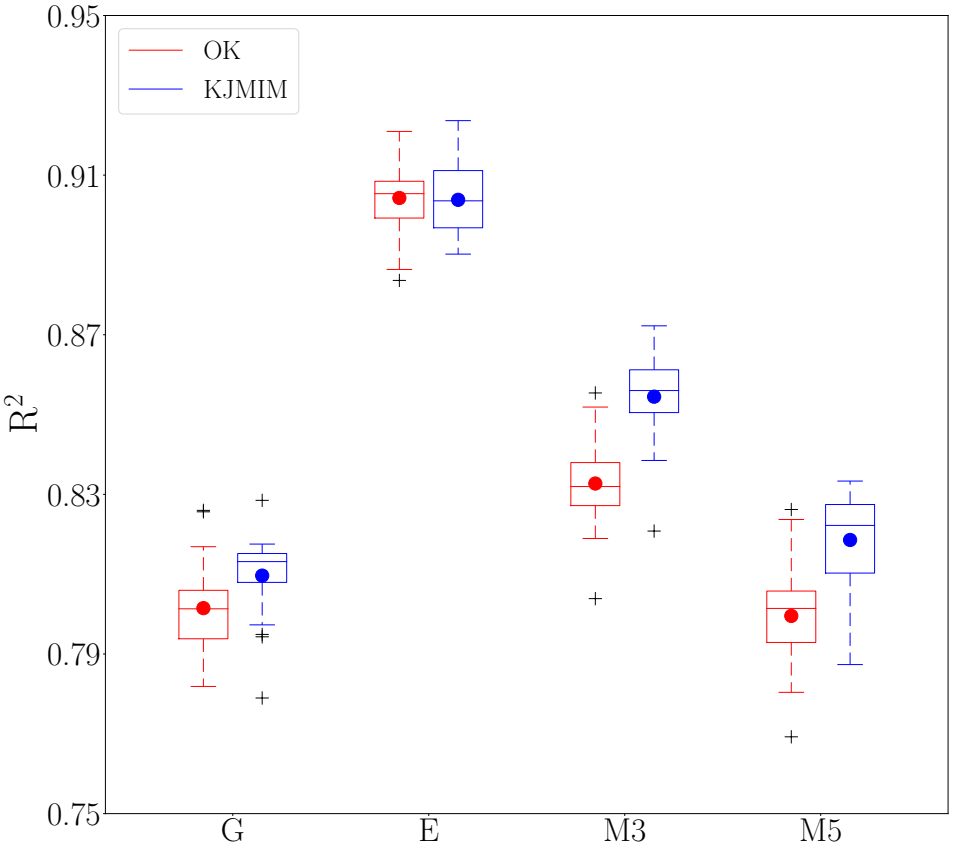
- Good modeling ability of the exponential and Matérn 3/2 kernels with the ordinary kriging (OK) model.
- Significant improvement in model performance with the feature selection-based method across all kernels.

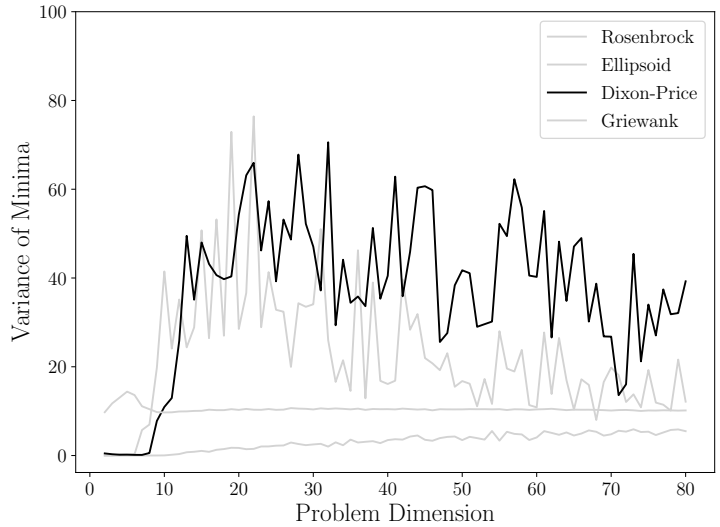




It is also important to carefully select the kernel for the 30-D Dixon-Price problem

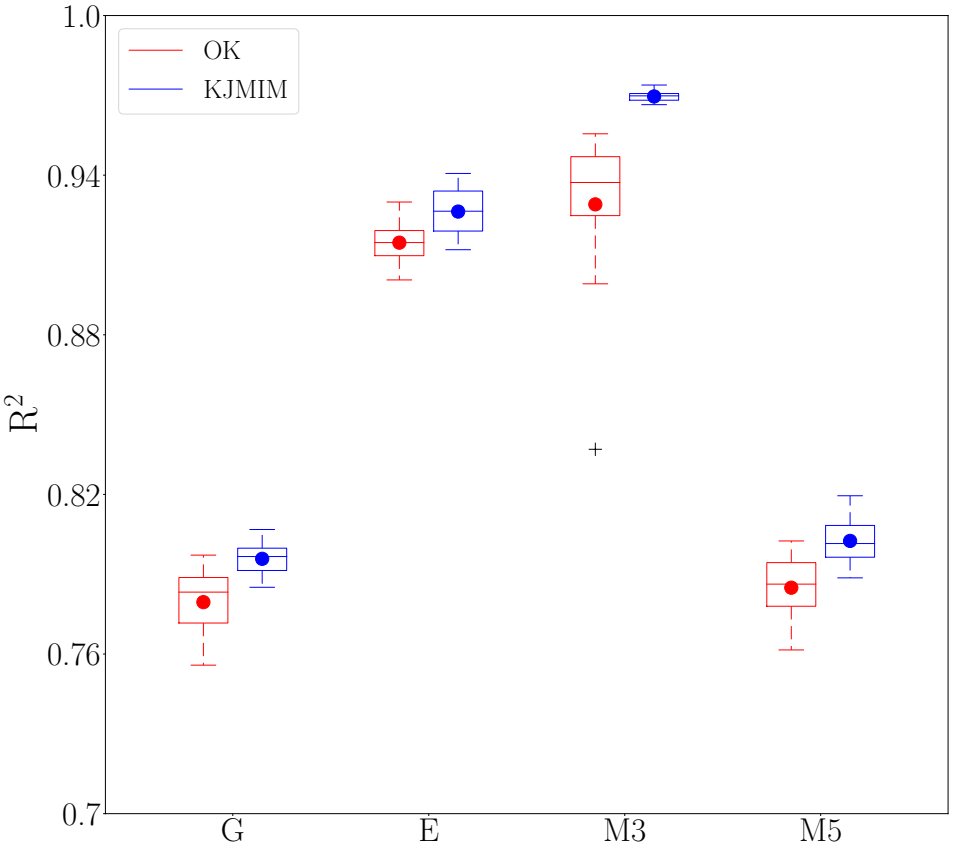
- Slight increase in performance is observed when the reduced models are used.
- The exponential kernel has great modeling capability.

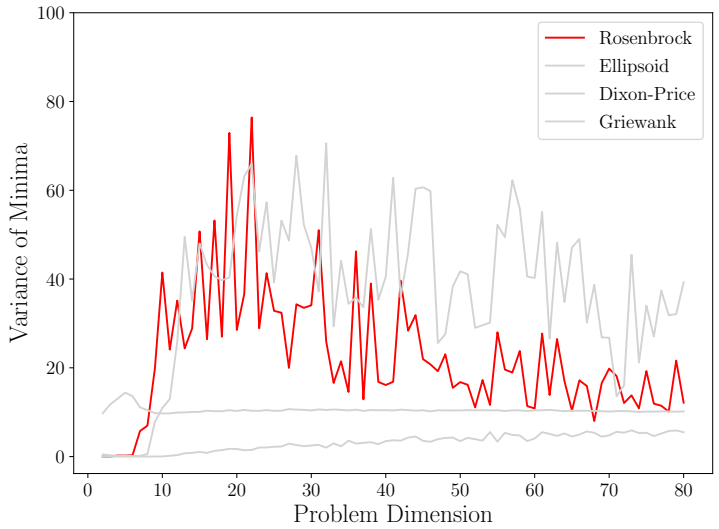




40-D Rosenbrock shows is multimodal and kernel selection is important

- Slight increase in performance is observed when the reduced models are used.
- The Matérn 3/2 and exponential kernels show good modeling performance with the multimodal problem.





The behaviour of perceived unimodal Rosenbrock function tend to change with an increase in dimension (Shang and Qiu, 2006)

Summary

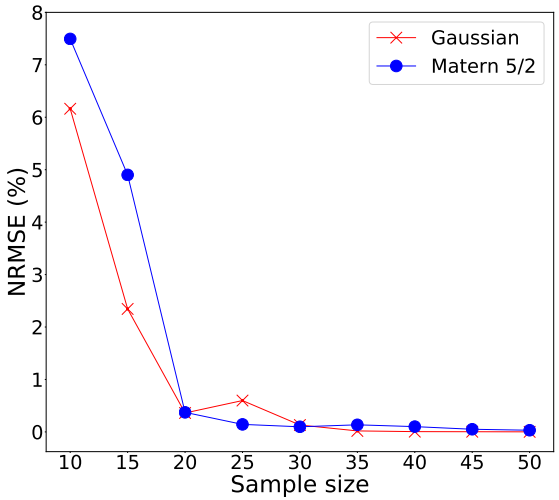
- When suitable kernels are not known, it is beneficial to use multiple kernel methods.
- Using different variable-kernel combinations (MIKL) can further improve model performance.
- Sensitivity to hyperparameter starting point is a major challenge for the Gaussian and Matérn 5/2 kernels.
- Techniques can be used to avoid poorly performing kernels during kernel selection.
- It is possible to improve the modeling accuracy of the the Gaussian kernel even in complex problems.
- Multimodality greatly affects the choice of kernels.
- The exponential and Matérn 3/2 kernels perform well with multimodal problems.

Section 3

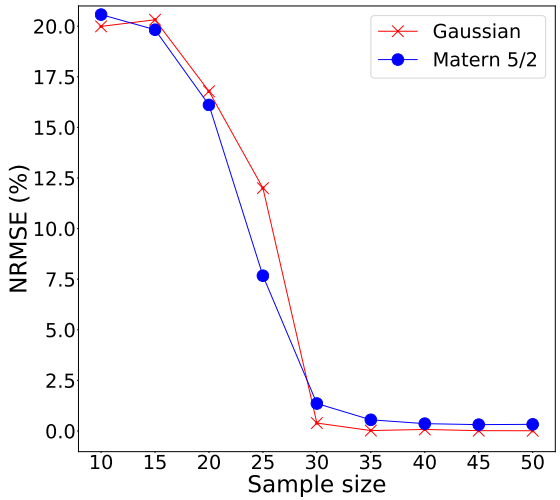
Choosing effective sample points

Adding more than enough points would not improve model performance

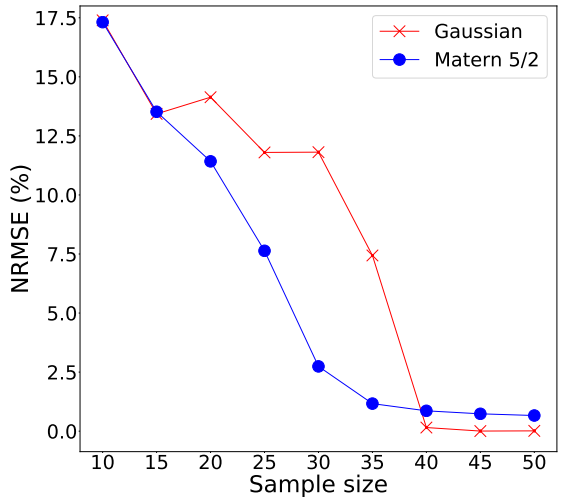
- More than required computational resources for model training.
- Kriging-based models have $\mathcal{O}(N_s^3)$ computational cost.
- It is important to know when to stop adding new points.



Branin



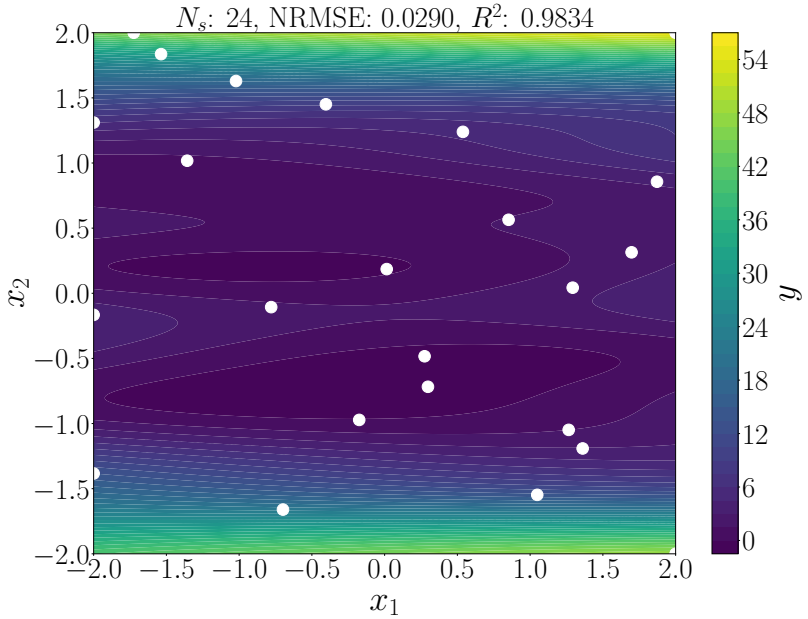
Himmelblau



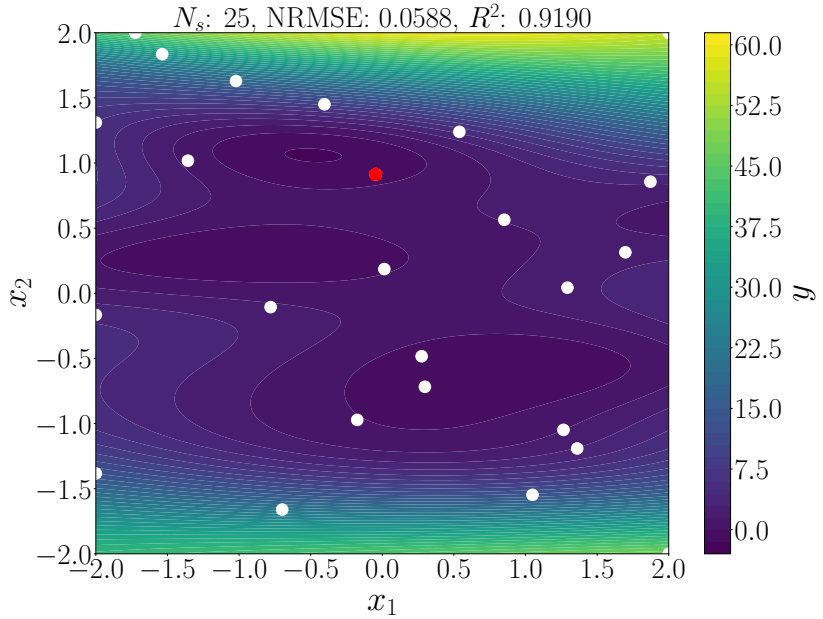
Camel back

Sometimes worse results are obtained with active learning

- Even with active learning strategies, care needs to be taken.
- It is of key importance to know where to stop the learning process.



24 points



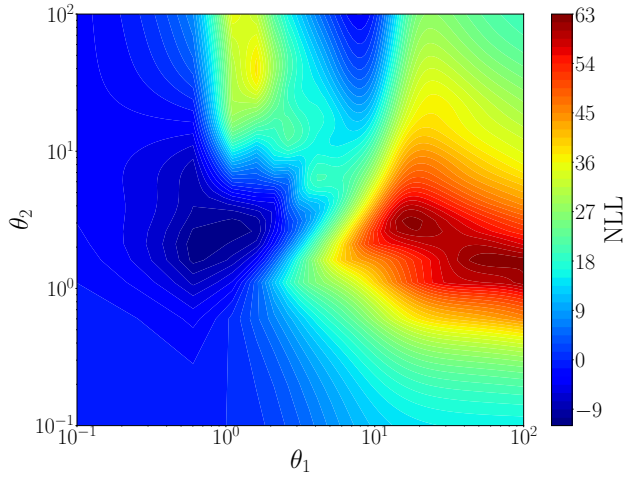
25 points

Camel back function

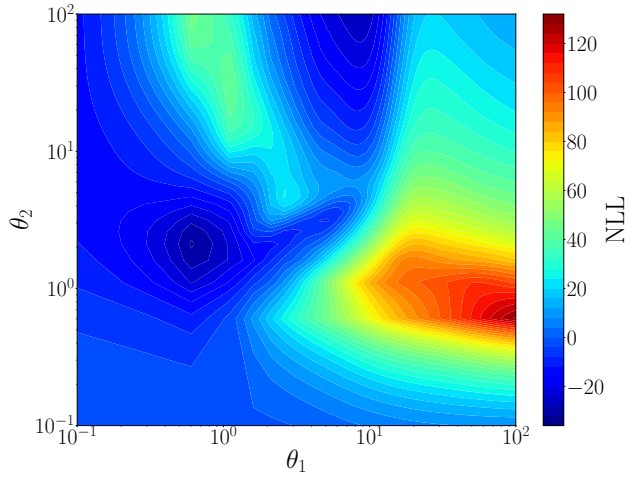
- Because kriging is an interpolation technique, it can drag the function to "overfit" the additional point.

A look into likelihood profile with increase in sample points with the camel back function

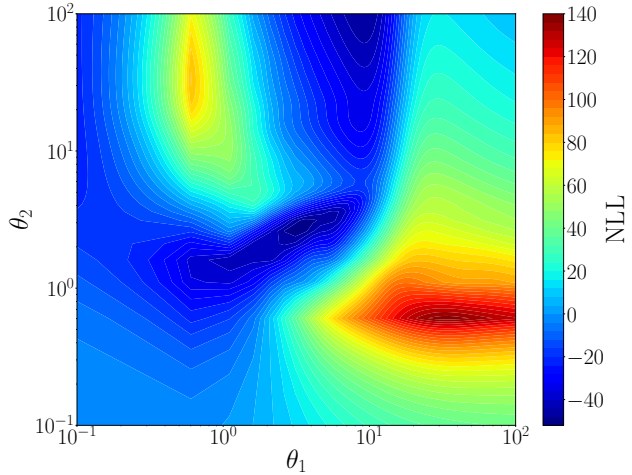
- Stability in likelihood profile can be reached with more points



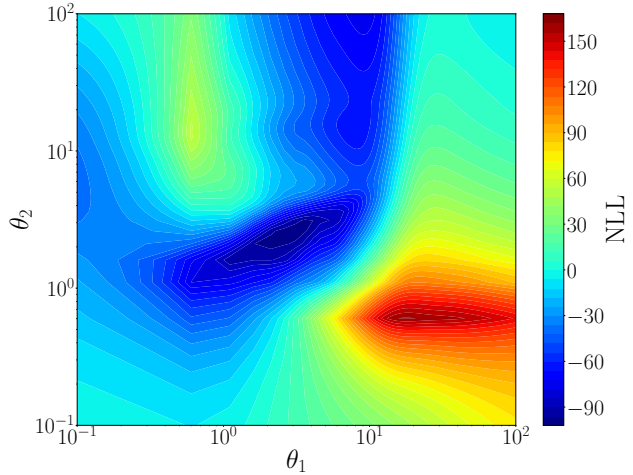
20 points



30 points

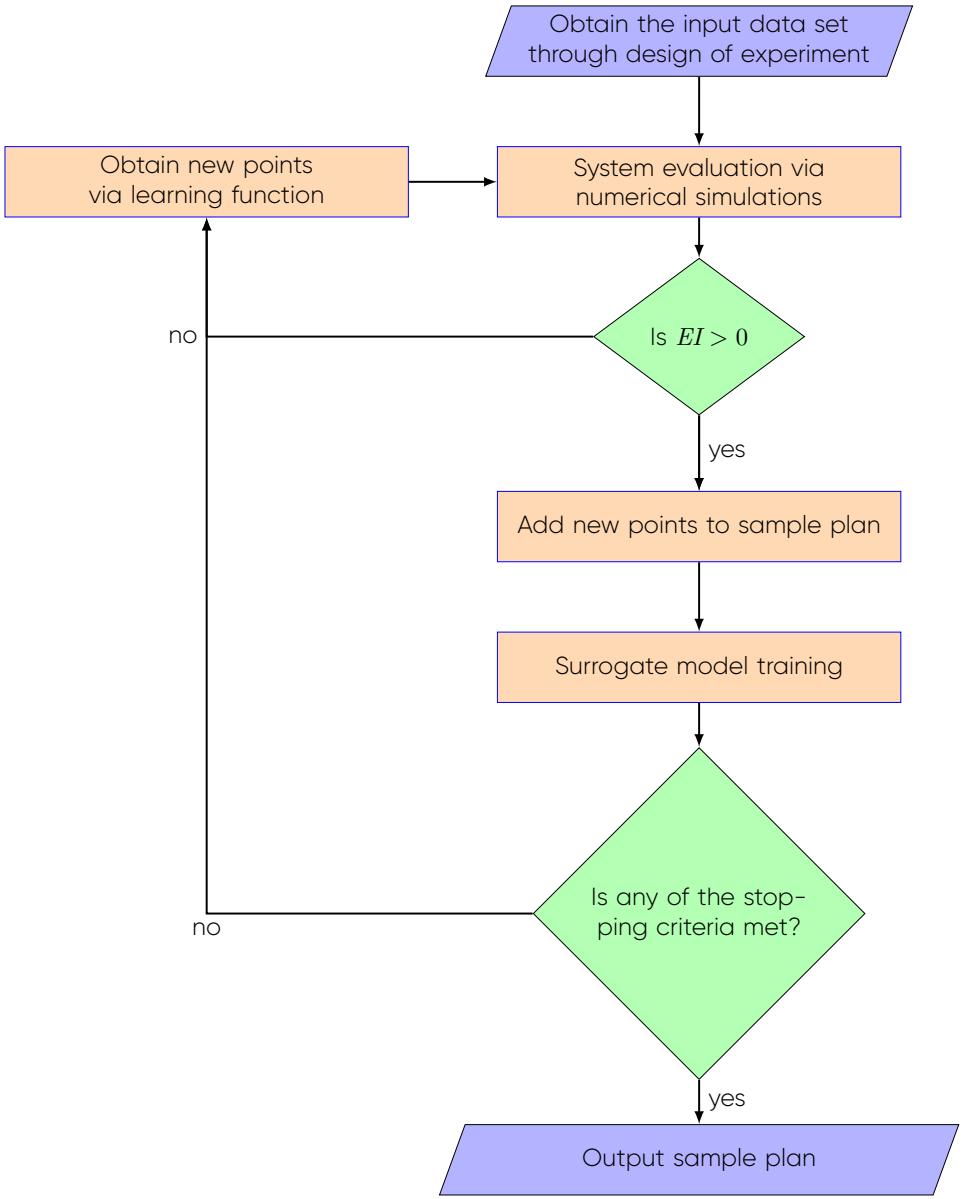


40 points



50 points

Proposed active learning framework



- We use learning function with **fixed exploitation-exploration weights**:
  $0.5\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 + 0.5\hat{s}(\mathbf{x})$
- We **introduce** the acquisition function, expected improvement (EI) to evaluate the size of improvement of candidate points: $a_{EI}(x) = \left(f_{min} - \hat{f}(\mathbf{x})\right)\Phi\left(\frac{f_{min}-\hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\phi\left(\frac{f_{min}-\hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)$

$\Phi(.)$: cumulative distribution function
$\phi(.)$: probability distribution function

$\hat{s}$: variance

Active learning with robust stopping criteria

Multiple stopping criteria-based active learning strategy is proposed.

- Likelihood-based criterion

$$\text{Absolute change in likelihood, } e_t = \frac{|LL_{t+1} - LL_t|}{LL_t} \leq \epsilon$$

- Successive iterations - ensure progressive improvement for n_e iterations

$$e_t > e_{t+1}$$

- Maximum stall criterion - No expected improvement (EI)

$$c > c_{max}$$

- Computational resource limitation - Maximum number of iterations

$$t > T_{max}$$

t: current iteration, c: number of failed expected improvement attempts, LL: Log-likelihood estimate

Benchmark method

1. For benchmarking with our method, we use the LOOCV-based stopping criteria technique.
2. We compute the leave-one-out estimate of the mean square error (MSE).

$$LOOCV_{N_s} = \frac{1}{N_s} \sum_{i=1}^{N_s} MSE_i.$$

3. We then obtain an approximation of the error, e

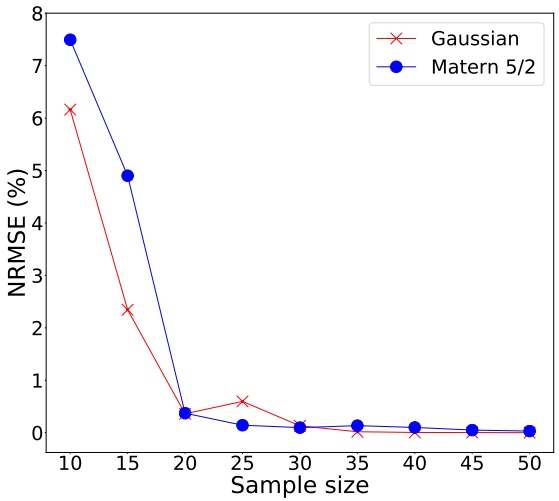
$$e_{N_s} = \sqrt{LOOCV_{N_s}}.$$

4. The stopping criterion term, α can be obtained by measuring the absolute change in error.

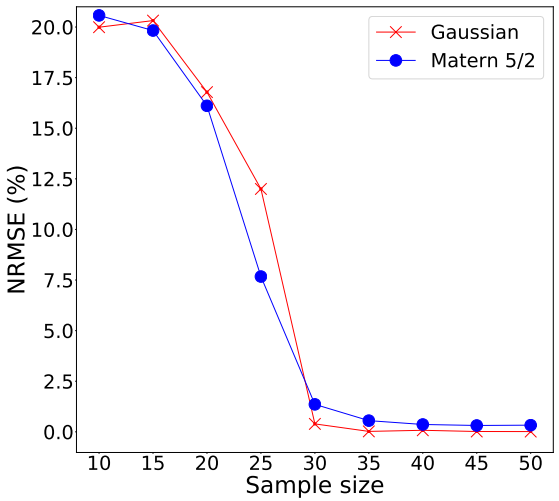
$$\alpha = \frac{|e_{N_s+1} - e_{N_s}|}{e_{N_s}},$$

A robust and precise convergence can be reached by using multiple criteria

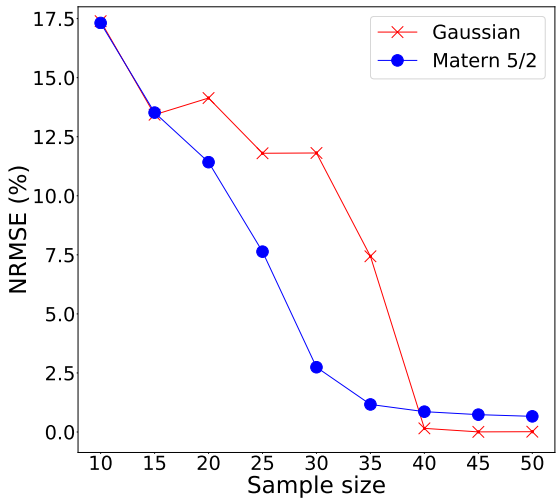
- With 20 initial sample points



Branin



Himmelblau



Camel back

Method	N_s^+	L_T (s)	R^2	N_s^+	L_T (s)	R^2
		Branin			Himmelblau	
AL_{MC}	13	17.99	0.9999	11	21.87	0.9998
$AL_{\alpha=0.01}$	11	102.15	0.9999	21	181.58	0.9999
$AL_{\alpha=0.05}$	10	94.47	0.9994	2	22.33	0.8219
		Camel back			Ackley	
AL_{MC}	19	19.65	0.9977	128	288.05	0.9077
$AL_{\alpha=0.01}$	4	28.62	0.9154	55	15133.75	0.8429
$AL_{\alpha=0.05}$	4	27.04	0.9130	8	1420.70	0.7664

N_s^+: Number of added points, L_T: Learning time, R^2: R^2 score

A robust and precise convergence can be reached by using multiple criteria

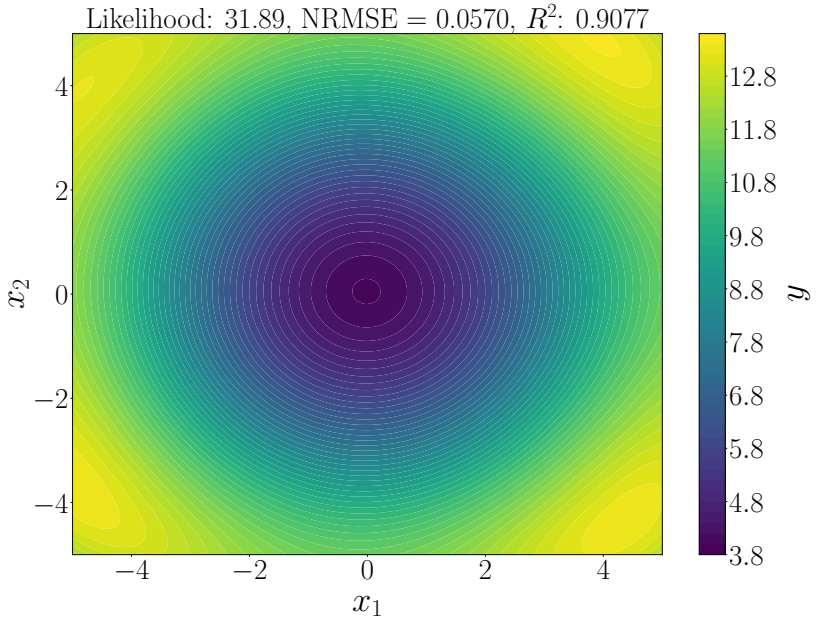
1. In almost all the benchmark cases, a good model accuracy is recorded when the learned function is used for model training.
2. The learning process by the AL_{MC} is completed in lesser time.

Method	N_s^+	L_T (s)	R^2	N_s^+	L_T (s)	R^2
		Branin			Himmelblau	
AL_{MC}	13	17.99	0.9999	11	21.87	0.9998
$AL_{\alpha=0.01}$	11	102.15	0.9999	21	181.58	0.9999
$AL_{\alpha=0.05}$	10	94.47	0.9994	2	22.33	0.8219
		Camel back			Ackley	
AL_{MC}	19	19.65	0.9977	128	288.05	0.9077
$AL_{\alpha=0.01}$	4	28.62	0.9154	55	15133.75	0.8429
$AL_{\alpha=0.05}$	4	27.04	0.9130	8	1420.70	0.7664

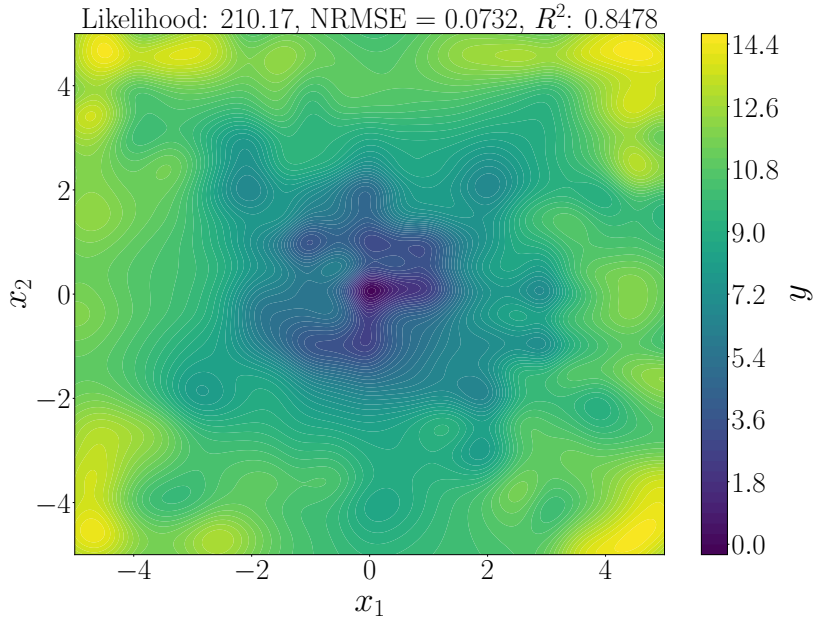
N_s^+: Number of added points, L_T: Learning time, R^2: R^2 score

Kernel selection becomes of importance with the learned Ackley function

Even worse model accuracy is recorded with the Matérn 5/2 kernel.



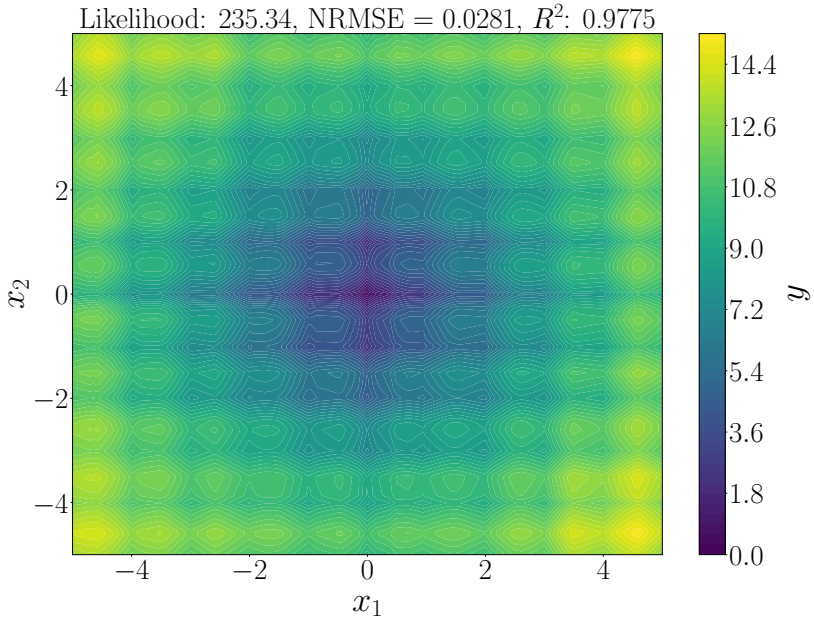
Gaussian



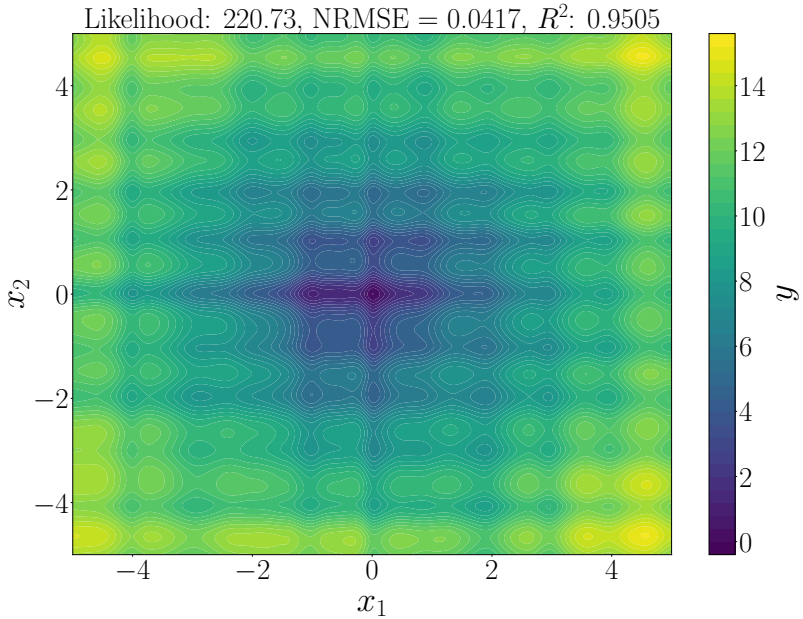
Matérn 5/2

Kernel selection becomes of importance with the learned Ackley function

Clear improvement in model accuracy are seen in the exponential and Matérn 3/2 kernels.



Exponential



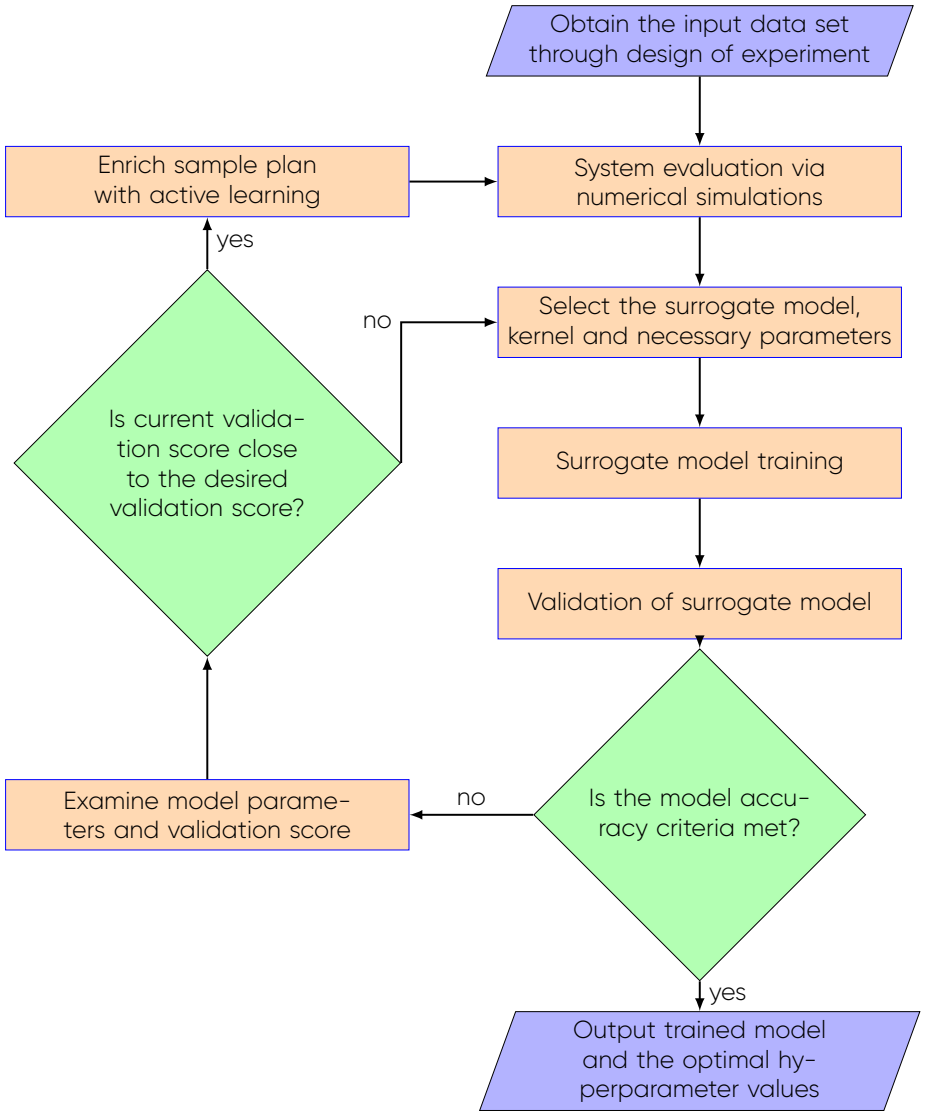
Matérn 3/2

Summary

- A robust active learning strategy should have multiple stopping criteria.
- Stability in the likelihood profile can be used in defining an efficient stopping criterion.
- Improved model performance can be obtained when effective kernel selections is combined with active learning strategies.

Conclusion

- Model structure should be preserved when reduced models are to be used within kriging framework.
- Kernel selections should be considered during surrogate-aided design process.
- Matérn 3/2 kernel should be used as the first choice in surrogates.
- With highly multimodal problems, the exponential kernel should be used for building the models.
- Multiple stopping criteria should be used within active learning strategies.



Original contributions

In section 1, we improved the **scalability** of kriging in high-dimensions

- showed that loss in model accuracy with existing methods is as a result of changes in model structure.
- proposed the use of joint mutual information maximization in preserving the model structure of kriging models.

In section 2, we improved kernel selections in kriging

- showed the benefits of using MIKL method when suitable kernels are not known.
- developed the optimal-ν and optimal-κ methods to automate kernel selections.
- showed that the exponential and Matérn 3/2 kernels are best suited for multimodal problems.

In section 3, we further improved sample selections

- introduced the use of expected improvement within active learning framework.
- developed robust and efficient multiple stopping criteria to improve the quality of sample points.

Achievements

- **Kehinde Sikirulai Oyetunde**, Rhea Patricia Liem. "Navigating kernel selections in kernel-based methods: The issues and possible solutions". *AIAA SciTech 2022, San Diego, CA.*
- Rhea Patricia Liem, **Kehinde Sikirulai Oyetunde**, Pramudita Satria Palar, and Koji Shimoyama. "Kriging with mixed kernel (MK) for complex aerospace problems". *Sixteenth International Conference on Flow Dynamics 2019*.
- **Kehinde Sikirulai Oyetunde**, Rhea Patricia Liem. "Efficient Kriging-based Modeling and Kernel Selections for High-dimensional Problems." (Prepared manuscript)
- **Kehinde Sikirulai Oyetunde**, Rhea Patricia Liem. "Robust Active Learning Technique with Efficient Multiple Stopping Criteria." (Prepared manuscript)

Awards

- Finalist Award, TECO Green Tech. International Contest, Taiwan '19
- Finalist Award, Rolls-Royce Data Innovation Challenge (RRDIC) Singapore '19

Appreciation

1. Advisor: Prof. Rhea P. Liem, for the continuous patience, mentoring and encouragement.
2. Committee members: Prof. K.Y. Michael Wong, Prof. Zhiwen Zhang, Prof. Wenjing Ye, Prof. Yanglong Lu.
3. Research Grants Council, Hong Kong, for funding my Ph.D. studies through HKPFS.
4. All my colleagues from the OCTAD Lab.
5. My friends here in Hong Kong.
6. Association of Nigerian Scholars in Hong Kong (ANSHK).
7. My family for their love and support.

Section 4

Appendix

Hyperparameter optimization

Hyperparameter optimization refers to **tuning** in surrogate modelling.

- It involves the minimization of a predefined loss function to obtain a set of optimal hyperparameters.
- The choice of optimizer is crucial to obtaining the required solution.

Some optimization algorithms

1. Constrained Nelder-Mead algorithm – simplex method
2. Sequential Least Squares Programming (SLSQP) – SQP method
3. Covariance Matrix Adaptation Evolution Strategy – evolutionary algorithm

Matérn kernels

The Matérn kernel has a functional form which can be expressed as;

$$\kappa(h, \theta, v) = \frac{1}{2^{v-1}\Gamma(v)} \left(2\sqrt{v}\frac{|h|}{\theta}\right)^{v} K_v \left(2\sqrt{v}\frac{|h|}{\theta}\right)$$

where $v \geq 1/2$ is the shape parameter, Γ is the Gamma function, and K_v is the modified Bessel function of the second kind.

v	Kernel
$\frac{1}{2}$	Exponential
$\frac{3}{2}$	Matérn 3/2
$\frac{5}{2}$	Matérn 5/2
∞	Gaussian

Performance across dimensions

Dimensionality has two main repercussions – Jorg *et. al.* (2019)[30]

1. The *search* on the surrogate model becomes **costly**
2. Surrogate model *building procedure* itself becomes difficult

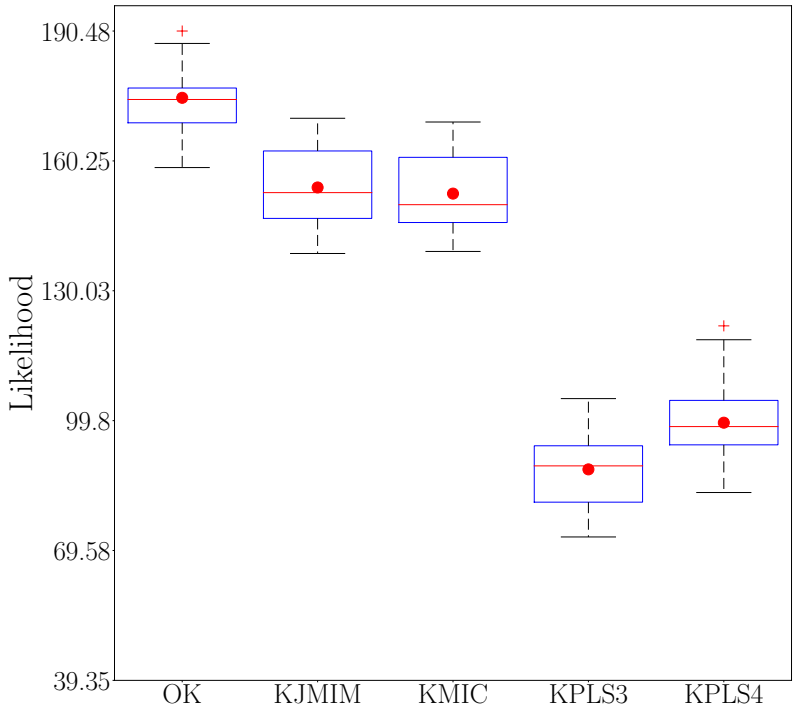
- n-dimensional cantilever beam problem

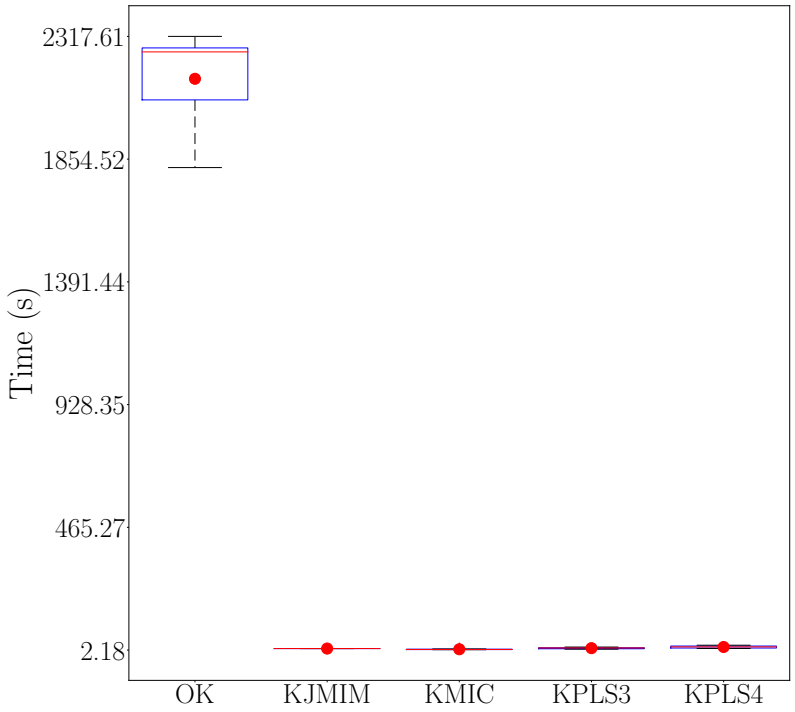
$$w = \frac{P}{3E}\sum_{i=1}^{n}\left[\frac{12}{bh^3}\left(\left(\sum_{j=i}^{n} l_j\right)^3 - \left(\sum_{j=i+1}^{n} l_j\right)^3\right)\right]$$

Input/Output	Quantity	Description
y_1	w	Tip deflection
x_i	$0.5 \leq l_i \leq 1.0$	Length of the i-th element
	b = 0.03	Width of the elements
	h = 0.50	Height of the elements
	P = 50 kN	Applied force at the tip
	E = 200 GPa	Young's modulus

[30] Open issues in surrogate-assisted optimization

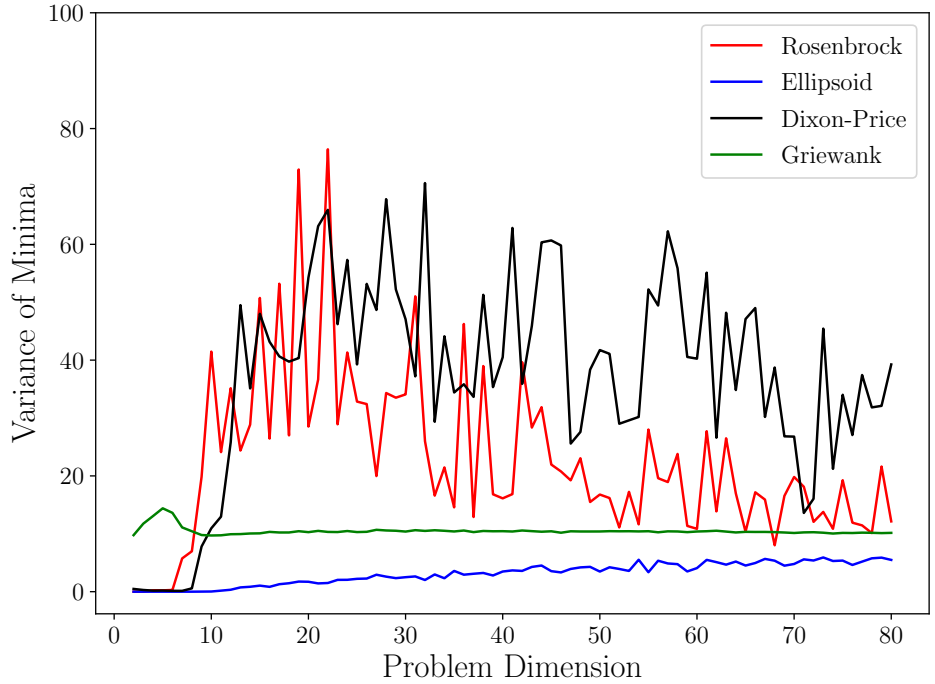
Likelihood and computational time benchmark for Griewank problem



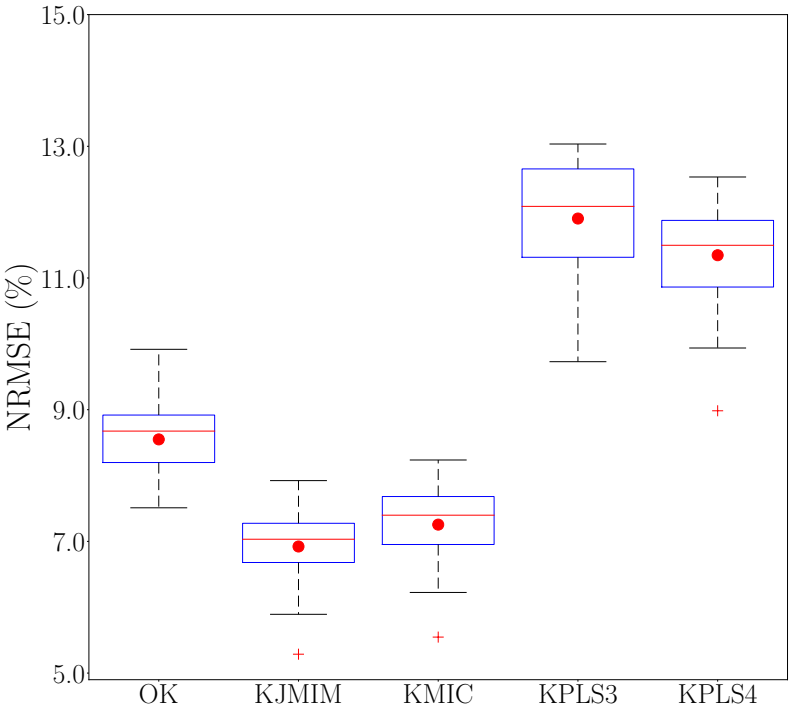


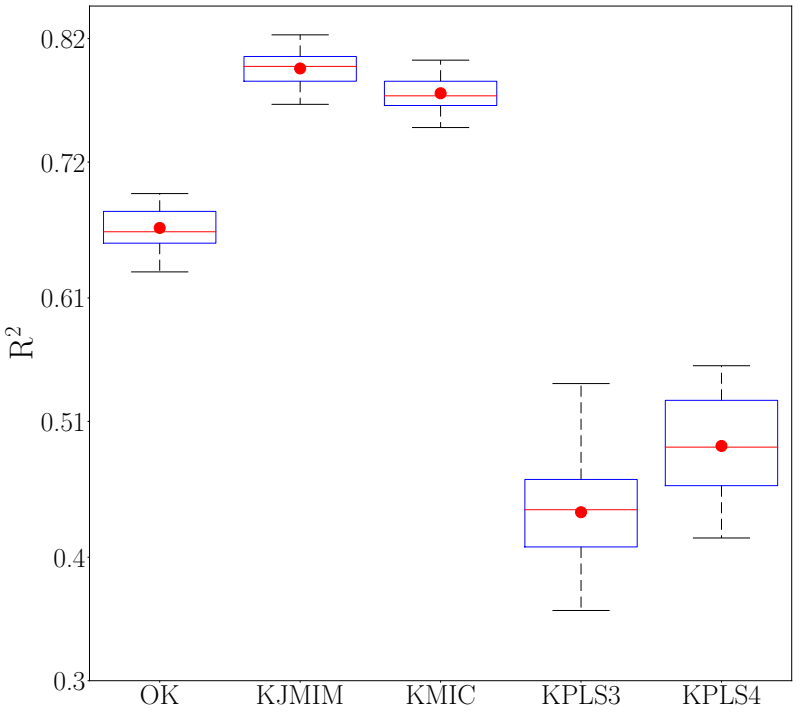
High-dimensional benchmark functions

Name	N_d	N_s	Expression
Ellipsoid	20	200	$f(x) = \sum_{i=1}^{20} ix_i^2, x_i \in [-5, 5], i = 1, ..., 20$
Dixon-Price	30	300	$f(x) = (x_1 - 1)^2 + \sum_{i=2}^{30} i(2x_i^2 - x_{i-1})^2, x_i \in [-10, 10], i = 1, ..., 30$
Rosenbrock	40	400	$f(x) = \sum_{i=1}^{39} \left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right], x_i \in [-5, 10], i = 1, ..., 40$
Griewank	80	500	$f(x) = \sum_{i=1}^{80} \frac{x_i^2}{4000} - \prod_{i=1}^{80} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1, x_i \in [-5, 5], i = 1, ..., 80$

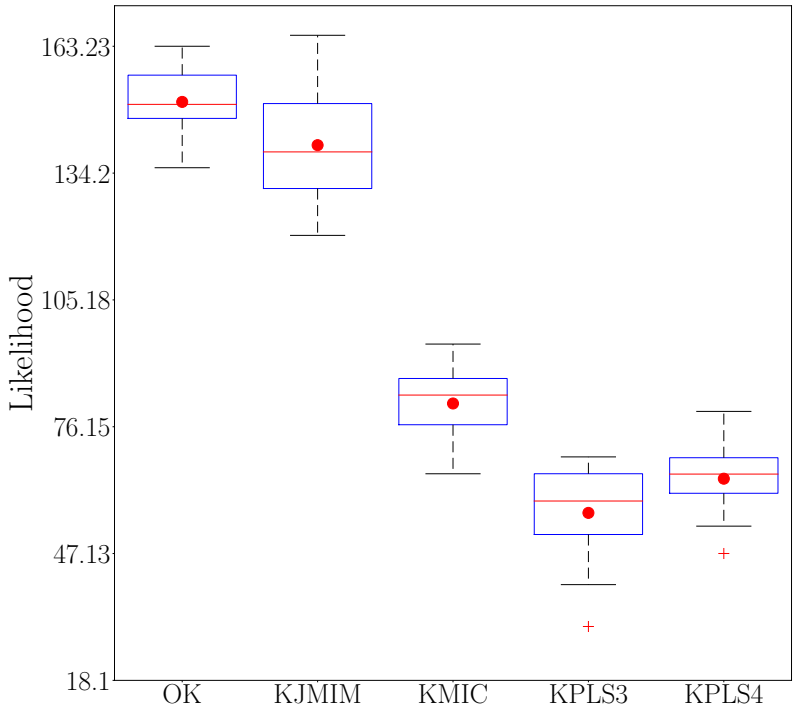


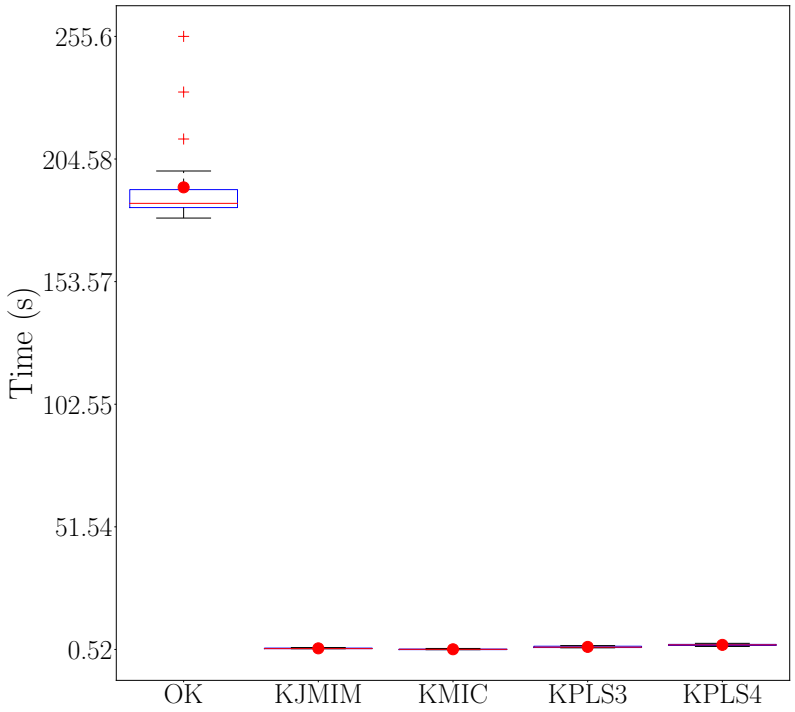
Model accuracy benchmark for Griewank problem



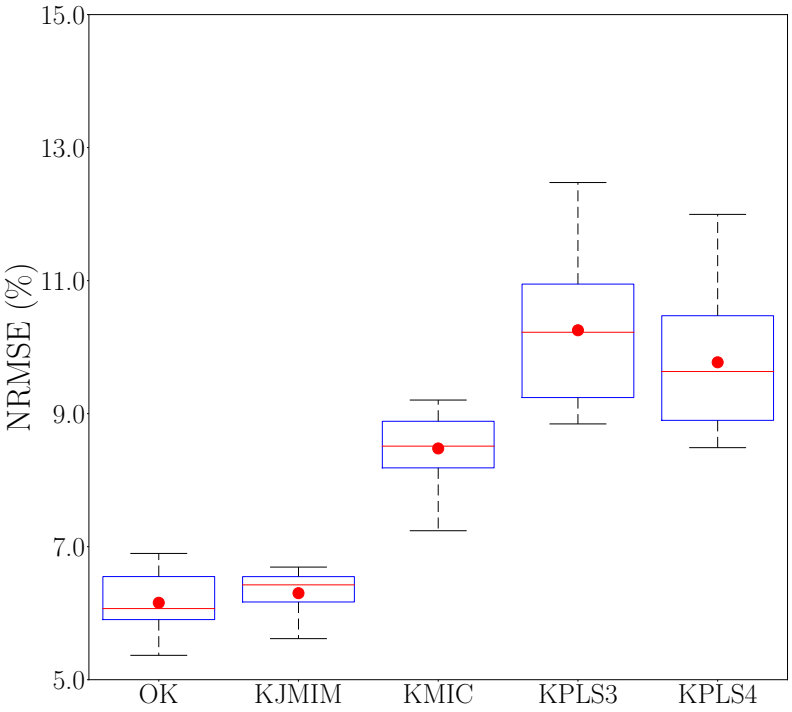


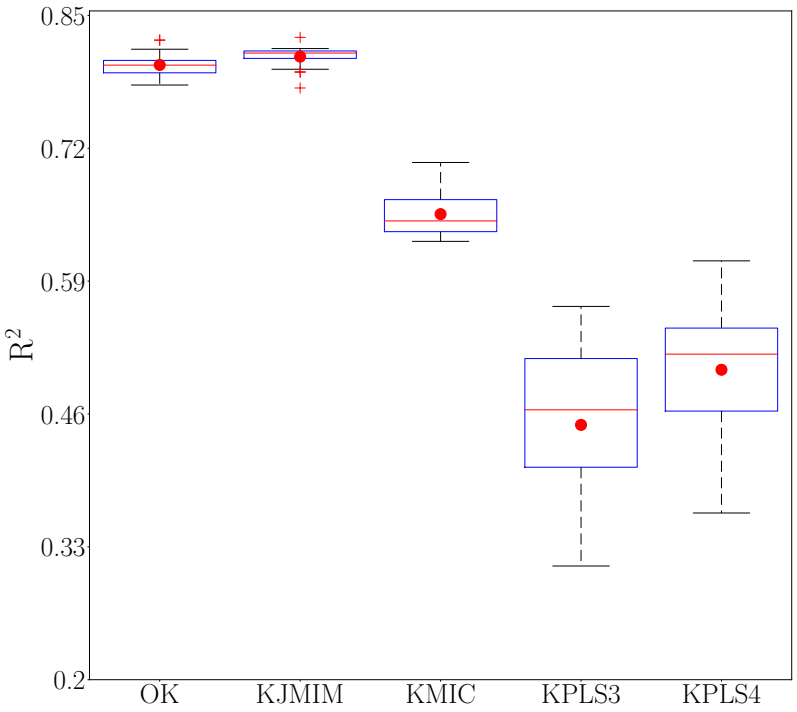
Likelihood and computational time benchmark for Dixon-Price problem



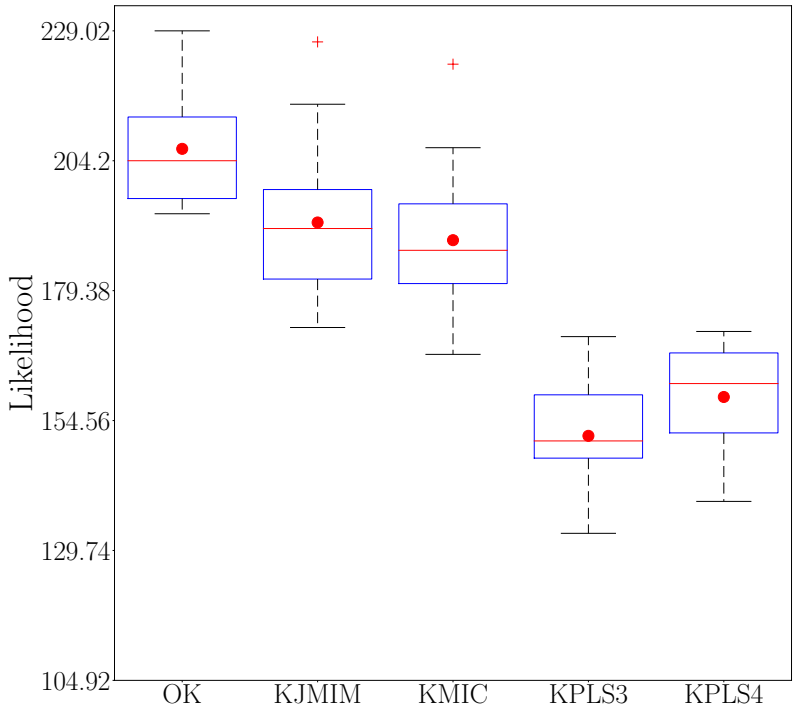


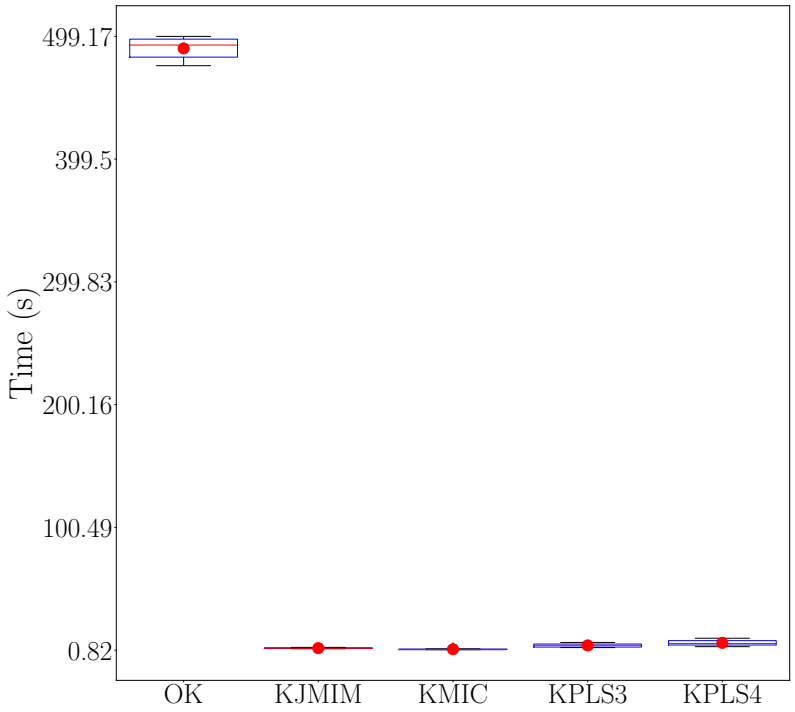
Model accuracy benchmark for Dixon-Price problem



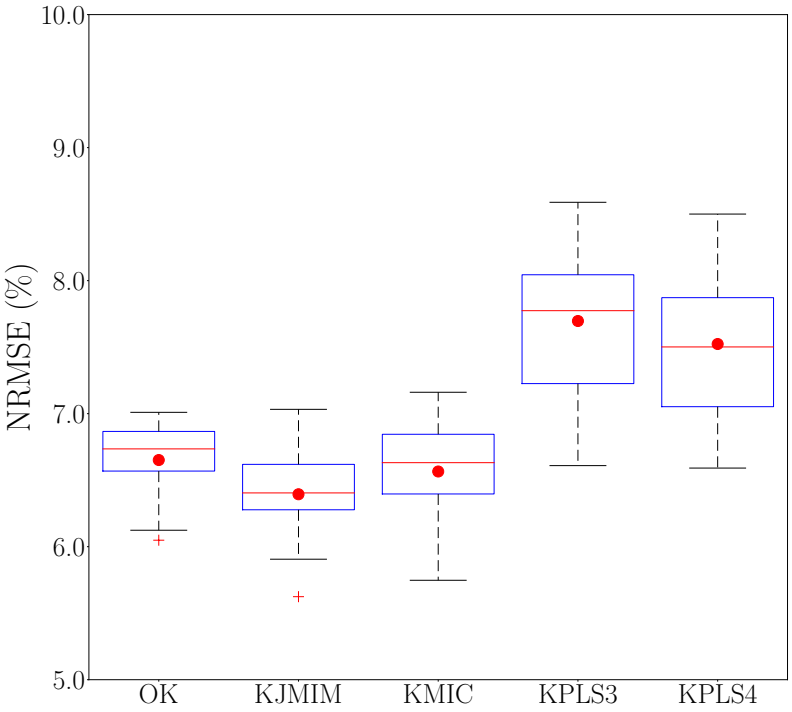


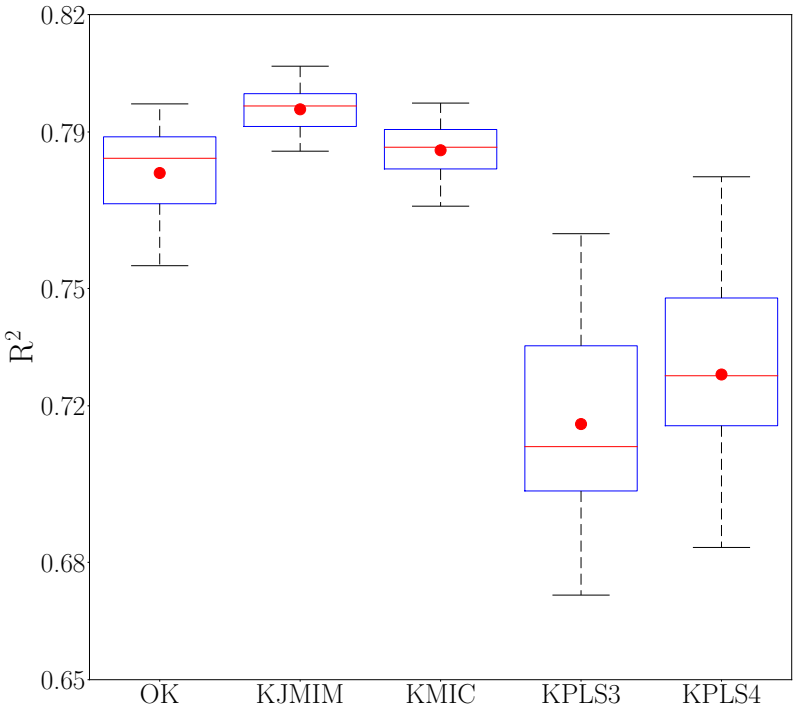
Likelihood and computational time benchmark for Rosenbrock problem



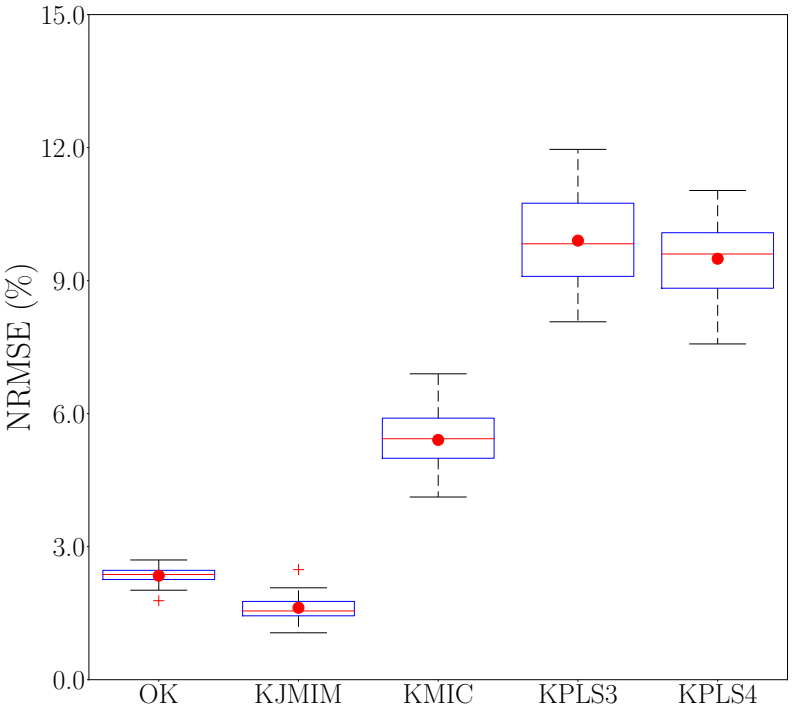


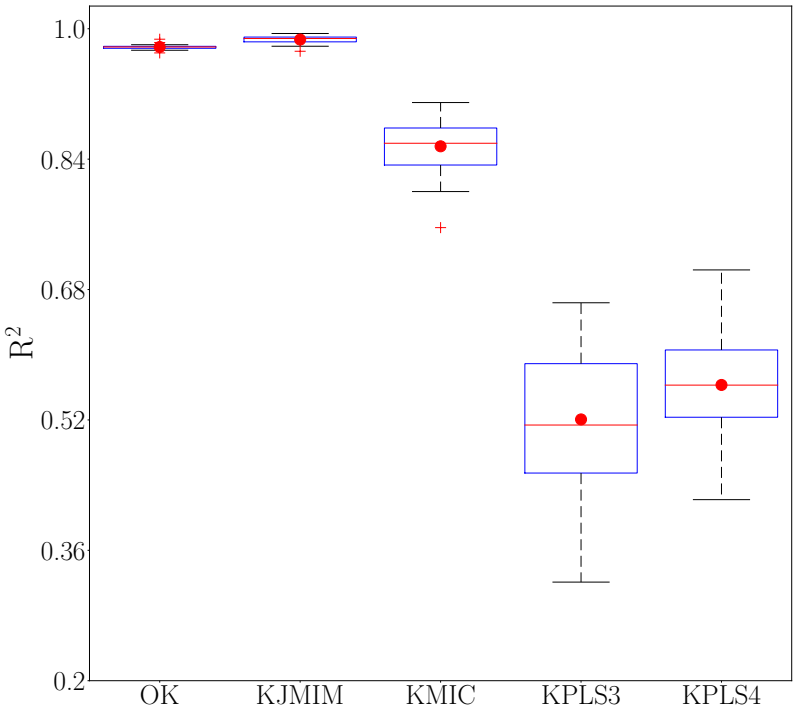
Model accuracy benchmark for Rosenbrock problem





Model accuracy benchmark – Ellipsoid problem





Summary of mean computational time and likelihood estimates

- **Computational time (s)**

Model	Ellipsoid (20-D)	Dixon-Price (30-D)	Rosenbrock (40-D)	Griewank (80-D)
OK	68.69	192.79	489.39	2157.67
KJMIM	0.29	0.97	2.41	7.93
KMIC	0.17	0.59	1.56	5.24
KPLS-3	0.89	1.59	4.61	9.45
KPLS-4	1.22	2.43	6.72	13.85

- **Likelihood estimates**

Model	Ellipsoid (20-D)	Dixon-Price (30-D)	Rosenbrock (40-D)	Griewank (80-D)
OK	156.33	150.49	206.48	174.94
KJMIM	150.45	140.60	192.41	154.07
KMIC	83.29	81.47	189.03	152.64
KPLS-3	43.16	56.43	151.64	88.46
KPLS-4	52.21	64.26	159.07	99.32

Ensemble Methods

- Proposed to overcome misspecification in optimization problems
- The goal of optimization is not to pursue global accuracy
- Ensemble techniques improved the robustness and performance of EGO
- Uncertainty vanishes (not useful in BO)
- Can either be local or global
- Uses weight to combine models
- Weight is computed
  - **`Akaike information criterion (AIC)`**
  - Bayesian information criterion (BIC)

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

EM formulations

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

The weight is computed;

$$w_i = \frac{exp(-0.5\Delta AIC_i)}{\sum_{j=1}^{K} exp(-0.5\Delta AIC_j)} \tag{2}$$

where,

$$AIC = -2\ln(L) + 2N_f \tag{3}$$

$$AIC_c = AIC + \frac{2N_f^2 + 2N_f}{n - N_f - 1} \tag{4}$$

$$\Delta AIC_i = AIC_i - AIC_{min} \tag{5}$$

$$N_f = len(\theta) + 2 \tag{6}$$

Composite kernel learning (CKL)

Current state-of-the-art method. Proposed by Palar *et. al.* (2019).

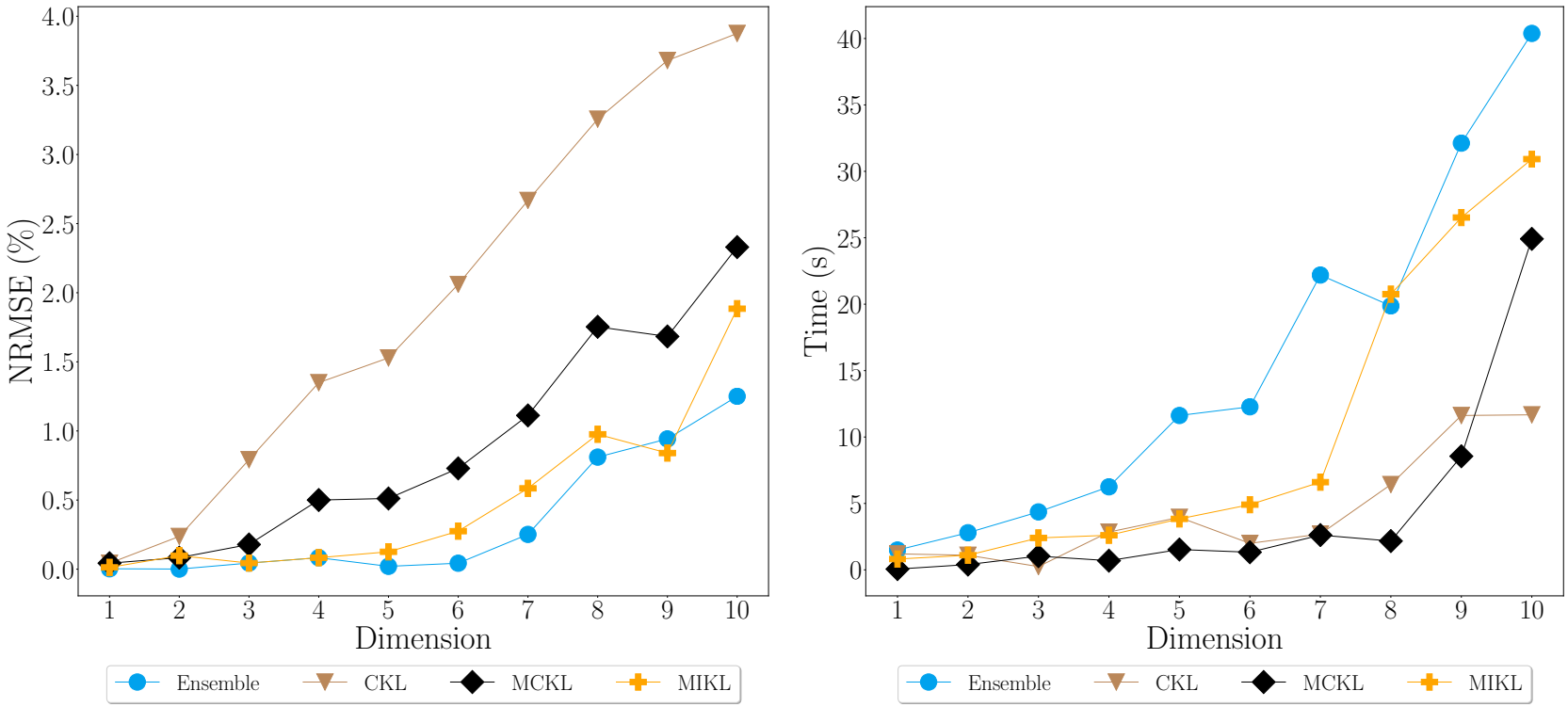
- It involves the construction of new kernels by combinations of existing kernels.
- Primarily focuses on the discovery of new kernels
- The new kernel is constructed so as to further optimize the likelihood function
- Uses weight to combine kernels
- Weight is usually optimized simultaneously with the hyperparameters.

$$R_{CKL}(h) = \sum_{i=1}^{K} w_i(x) R_i(h) \tag{7}$$

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

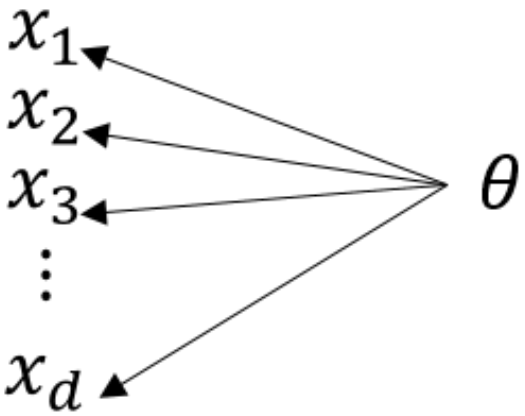
Multiple kernel methods might not be suitable for high-dimensional problems

- With an increase in problem dimension, there is a significant increase in training time.

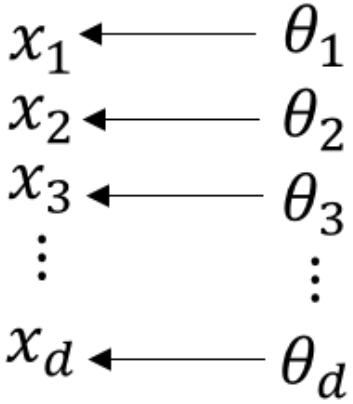


Cantilever beam problem (50 training points)

Isotropy/Anisotropy with kernel hyperparameters



Isotropy



Anisotropy

- Anisotropic correlation functions are advised when each variable has a distinct physical meaning (Hang and Steinwart, 2018).
- It gives more flexibility in modeling at the expense of a more complex maximum likelihood estimate (Currin *et. al.*, 1991).

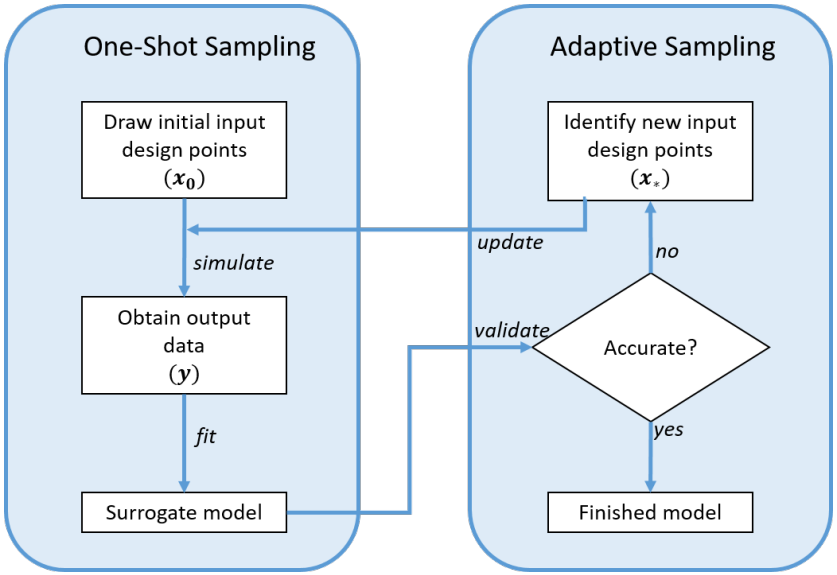
Axial Transonic Rotor Benchmark - NRMSE (%)

	Lower limit	Upper limit
Twist (radians)	-0.125	0.22
Sweep (radians)	-0.02	0.05

- The result for the Axial Transonic Rotor (total pressure ratio problem) shows our methods selected the better performing kernels - Gaussian and Matérn 3/2 kernels.
- For the adiabatic efficiency problem, our methods completely avoided all kernels except for the Gaussian kernel, which happens to be the only one with acceptable NRMSE value

	Total Pressure Ratio	Adiabatic Efficiency
Gaussian	0.7786	1.7655
Exponential	1.1650	2.2920
Matérn 3/2	0.6906	2.7961
Matérn 5/2	1.6718	3.6433
Optimal-ν method	0.7793(G)	1.7639(G)
Optimal-κ method	0.8067(M3)	1.6713(G)

Active learning strategy



Algorithm Typical active learning strategy

Input: Function, $f(.)$, Initial sample size, n_i
Output: $\mathbf{x}^{opt}$, $\mathbf{y}^{opt}$

Draw initial samples, $\mathbf{x_1}$
Obtain the functional value of the input data, $y_1 = f(\mathbf{x_1})$
Initialize iteration, $t = 1$
while stopping criterion is not met **do**
    Calculate GP posterior $q_t(f|S_t)$
    Obtain next proposed point by maximizing the learning function.
    Obtain functional value of new point: $\mathbf{y}_{t+1} = f(\mathbf{x}_{t+1})$
    Update dataset: $S_{t+1} \leftarrow S_t \cup \{(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\}$
    $t \leftarrow t + 1$
end while

Active learning with robust stopping criteria

Absolute change in likelihood, $e_t = \dfrac{|LL_{t+1} - LL_t|}{LL_t}$,

Algorithm Active learning with multiple stopping criteria

Input: Function, $f(.)$, Initial sample size, n_i, Threshold point, ϵ, maximum it
check c_{max}, Size of considered likelihood change, n_e
Output: $\mathbf{x}^{opt}$, $\mathbf{y}^{opt}$

Draw initial samples, $\mathbf{x_1}$
Obtain the functional value of the input data, $y_1 = f(\mathbf{x}_1)$
Initialize iteration, $t = 1$
while $t < T_{max}$ **do**
    Calculate GP posterior $q_t(f|S_t)$
    Select starting point within $\mathbf{x}$-space
    Obtain next proposed point by maximizing the learning function
    Compute the expected improvement of the proposed point, $EI(\mathbf{x}_{t+1}^{propos}$
    $c \leftarrow 0$
    while $EI(\mathbf{x}_{t+1}^{proposed}) = 0$ **OR** $c < c_{max}$ **do**
        Reject $\mathbf{x}_{t+1}^{proposed}$
        Select another starting point within $\mathbf{x}$-space
        Obtain next proposed point by maximizing the learning function
        $c \leftarrow c + 1$
    end while
    if $c \geq c_{max}$ **then**
        break
    else
        Accept $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1}^{proposed}$
    end if
    Obtain functional value of new point: $\mathbf{y}_{t+1} = f(\mathbf{x}_{t+1})$
    Update dataset: $S_{t+1} \leftarrow S_t \cup \{(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\}$
    Calculate GP posterior $q_{t+1}(f|S_{t+1})$
    Initialize E = { }
    Compute absolute change in likelihood, e_t
    $E \leftarrow E \cup e_t$
    if $e_t < \epsilon$ **then**
        for $j = t - n_e, t + 1 - n_e, ..., t$ **do**
            if $e_j < e_{j+1}$ **OR** $e_j > \epsilon$ **then**
                break
            end if
        end for
    end if
    $t \leftarrow t + 1$
end while

Importance of computing size of improvement of candidate points

- Rich samples are only added.
- Numerical instability issues associated with the correlation matrix can be avoided.
- It can be used to define a stopping point.

Learning functions

A refinement criterion (RC) is maximized to achieve learning

$$\mathbf{x}^{n+1} = \underset{\mathbf{x}^* \in \mathcal{X}}{\operatorname{argmax}}\, RC(\mathbf{x}^*).$$

- Expected Prediction Error (EPE)

$$EPE(\mathbf{x}) = \left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 + \hat{s}(\mathbf{x}).$$

- Maximizing Expected Prediction Error (MEPE)

$$EPE(\mathbf{x})^{\alpha} = \alpha \left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 + (1-\alpha)\hat{s}(\mathbf{x})$$

- Expected Improvement for Global Fit (EIGF)

$$RC_{EIGF}(\mathbf{x}) = \left(\hat{f}(\mathbf{x}) - f(\mathbf{x}^*)\right)^2 + \hat{s}(\mathbf{x}),$$

- Smart Sampling Algorithm (SSA)

Common acquisition functions

- Probability of improvement

$$a_{PI}(x) = \frac{1}{2}\left[1 + erf\left(\frac{f_{min} - \hat{f}(\mathbf{x})}{\hat{s}\sqrt{2}}\right)\right].$$

- Expected Improvement

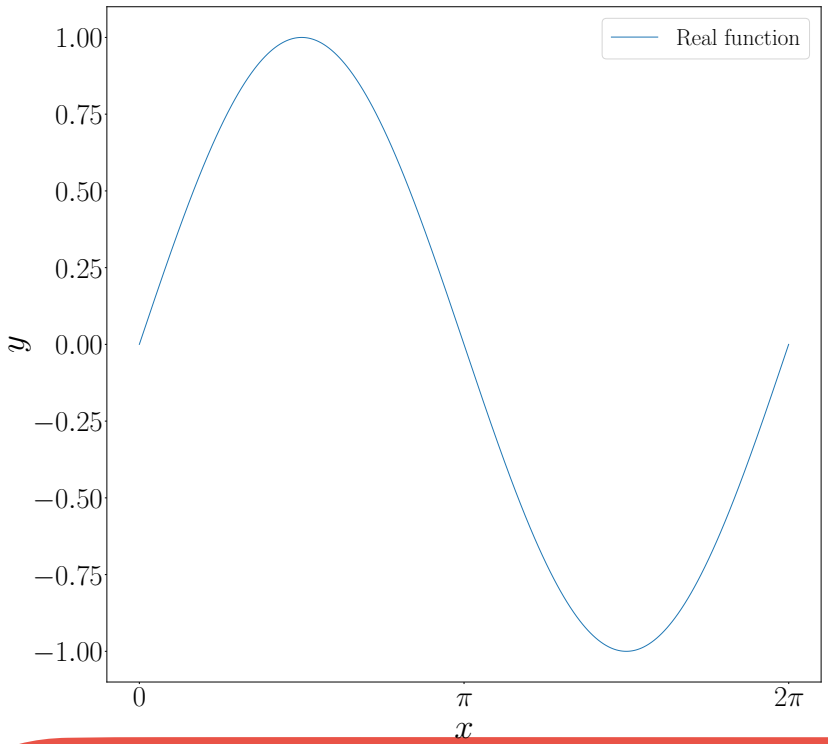
$$a_{EI}(x) = \left(f_{min} - \hat{f}(\mathbf{x})\right)\Phi\left(\frac{f_{min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\phi\left(\frac{f_{min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right),$$

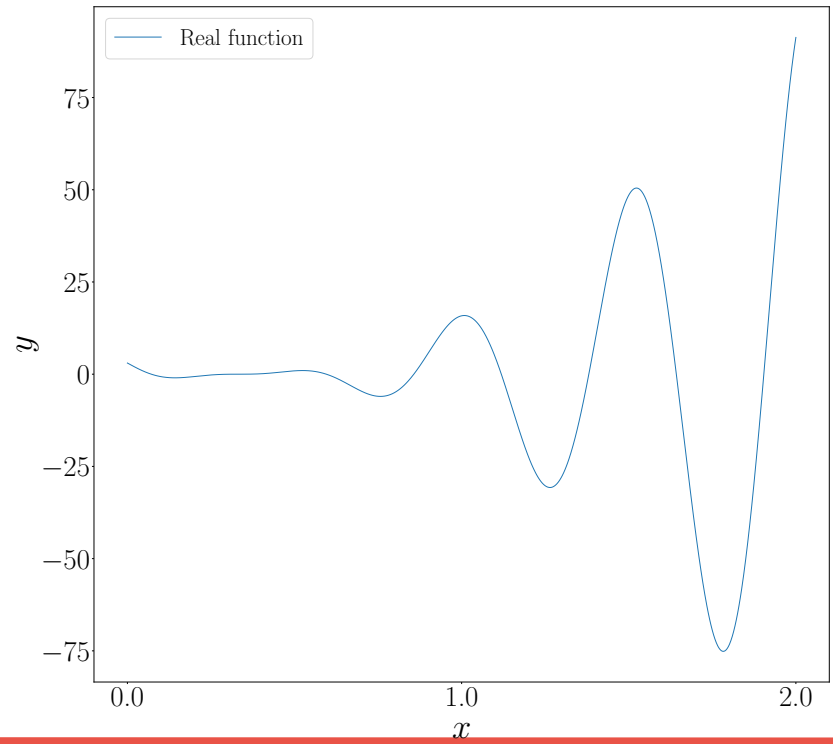
- Upper confidence bound

$$a_{UCB}(\mathbf{x}; \beta) = \hat{f}(\mathbf{x}) + \beta\sigma(\mathbf{x})$$

Is it really important to select good points for training?

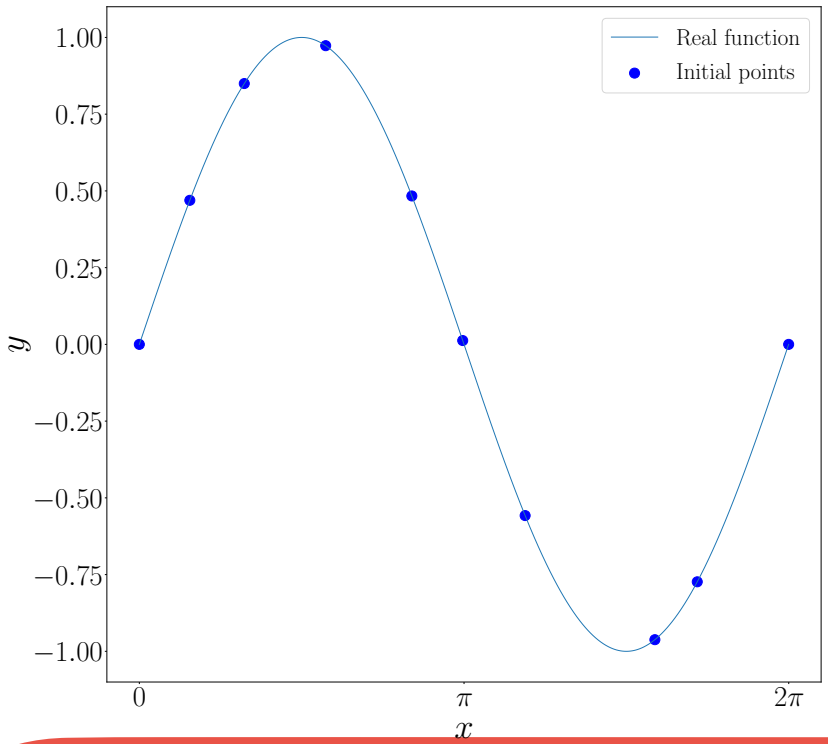
- Different numbers and positioning of sample points are required for different problems.
- Space-filling sampling method focus on the input space.
- The new points can be added *adaptively* (where needed) using active learning.

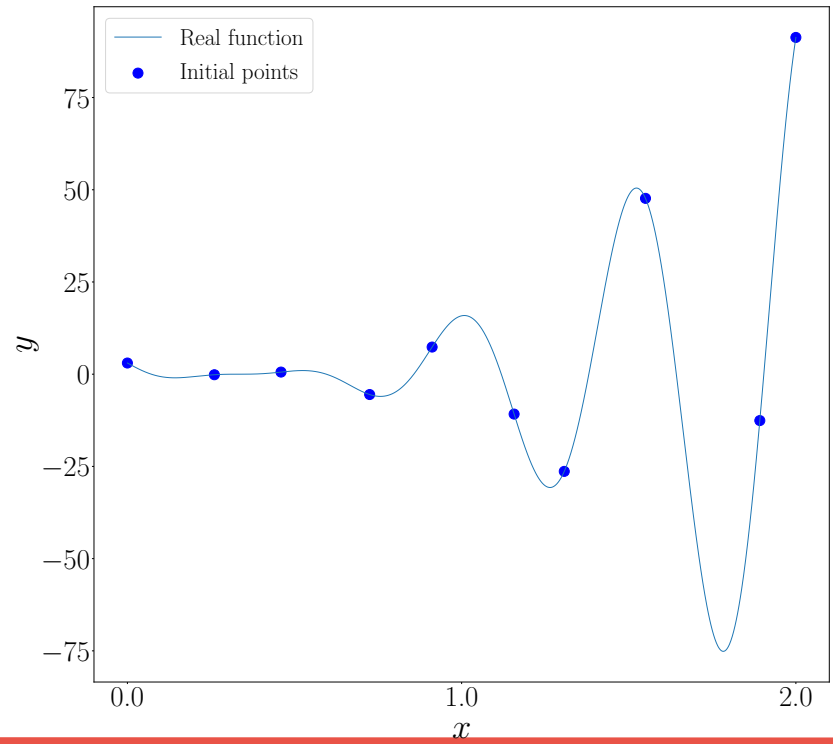




Is it really important to select good points for training?

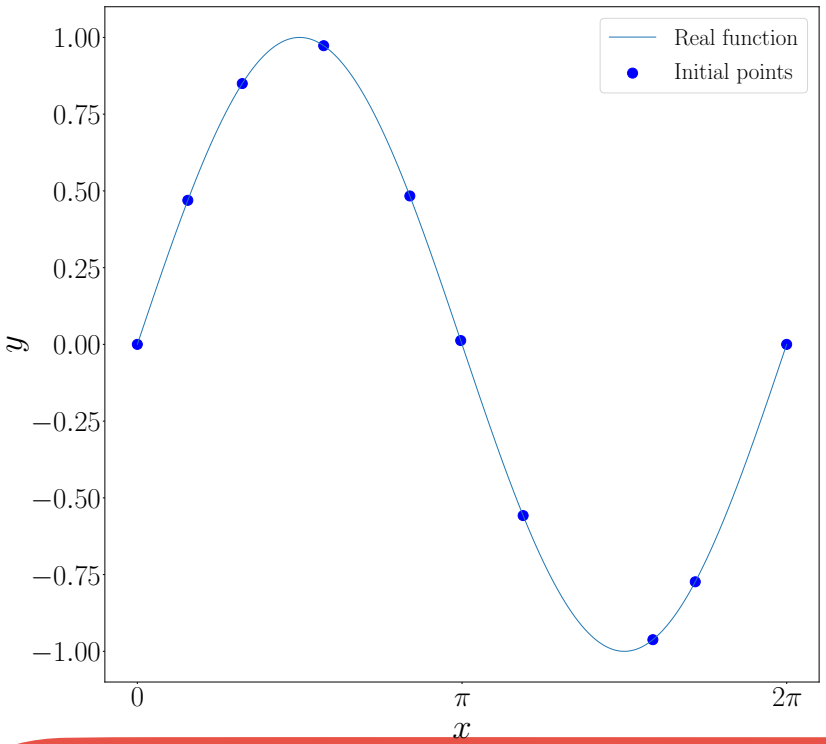
- Different numbers and positioning of sample points are required for different problems.
- Space-filling sampling method focus on the input space.
- The new points can be added *adaptively* (where needed) using active learning.

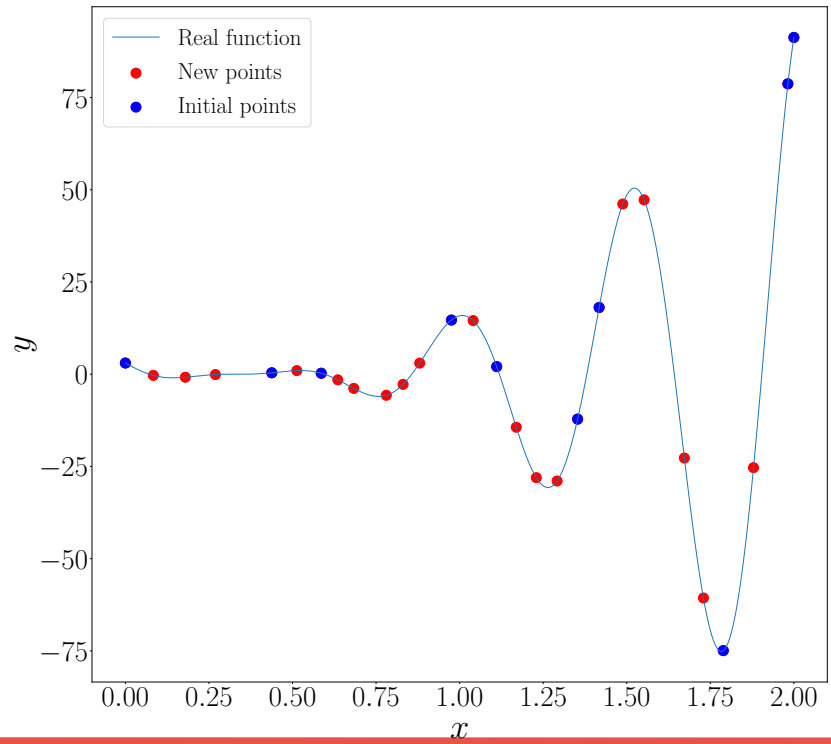




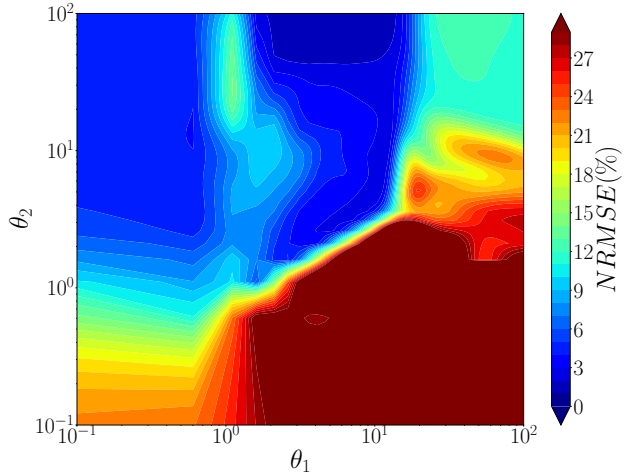
Is it really important to select good points for training?

- Different numbers and positioning of sample points are required for different problems.
- Space-filling sampling method focus on the input space.
- The new points can be added *adaptively* (where needed) using active learning.

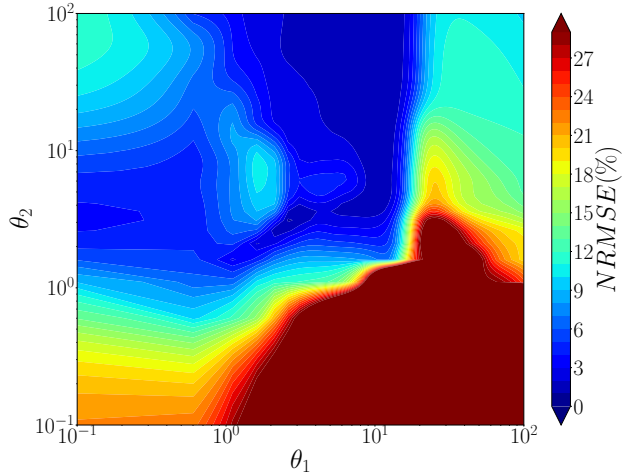




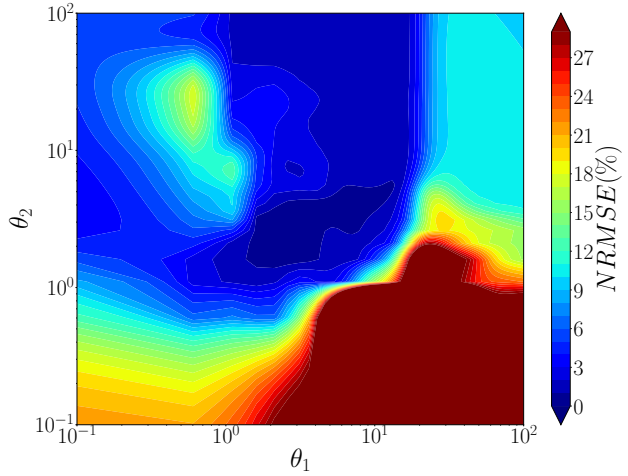
Changes in the model accuracy with increase in sample points



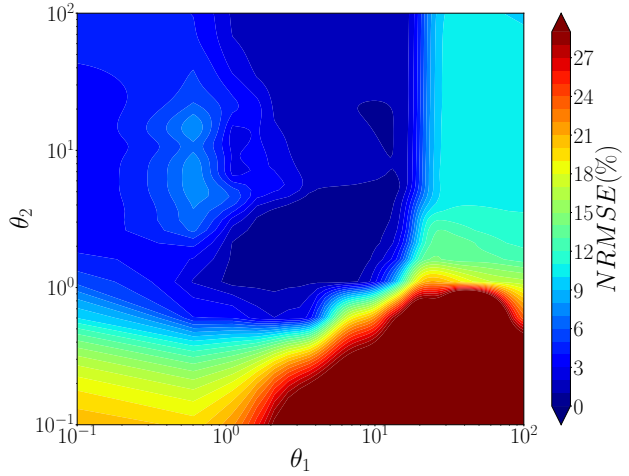
20 points



30 points

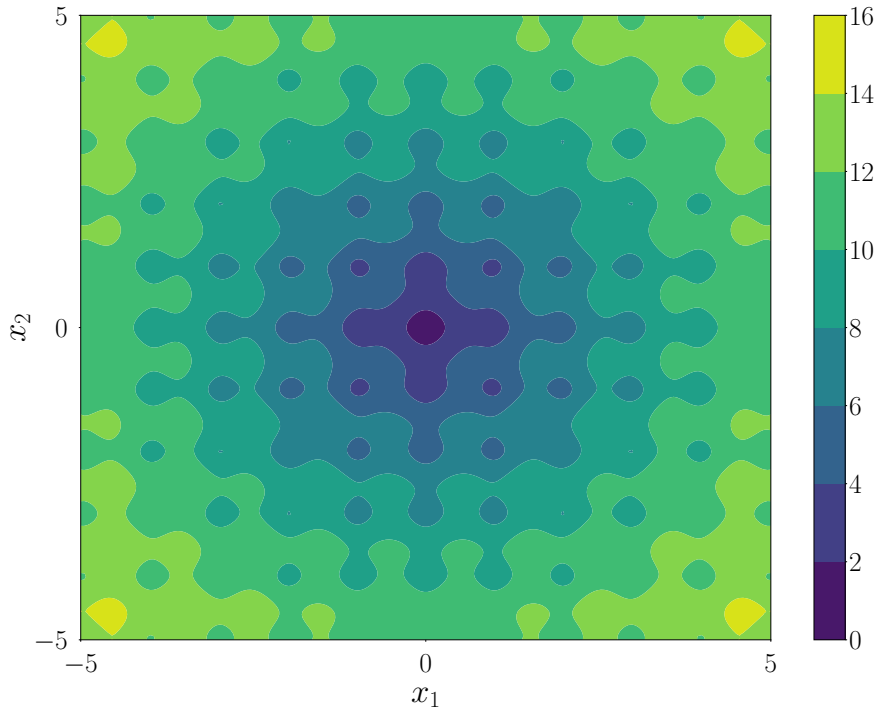


40 points

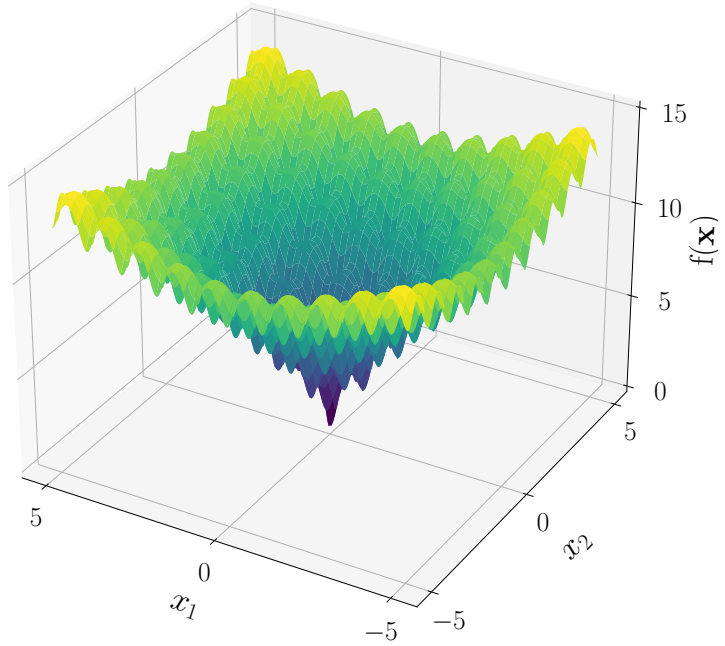


50 points

Ackley function



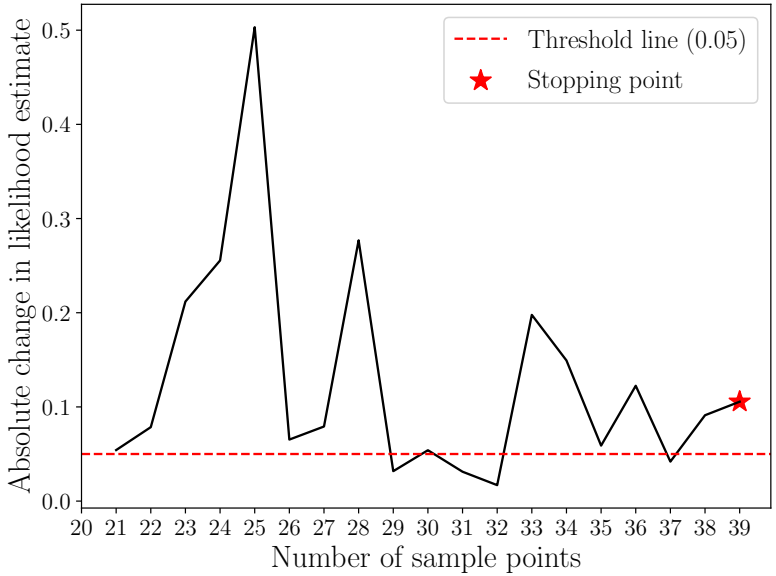
Contour profile



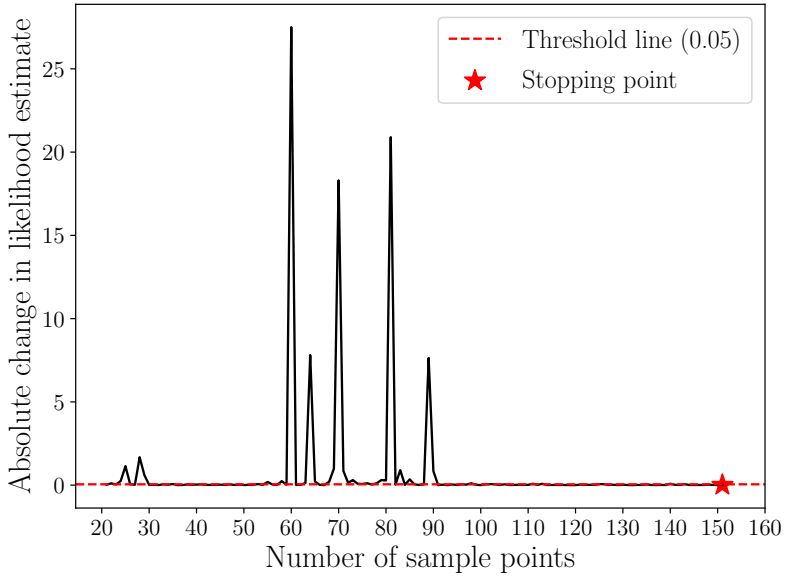
Response surface

A robust and precise convergence can be reached by using multiple criteria

- With 20 initial sample points



Camel back

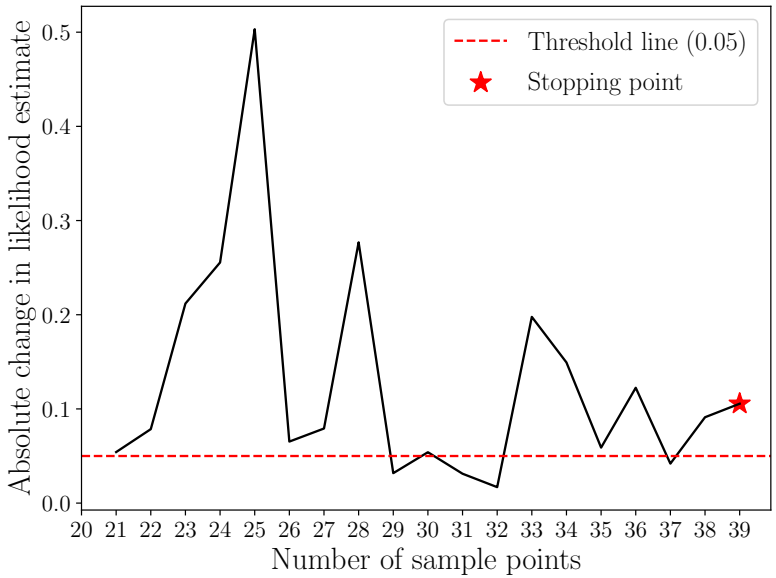


Ackley

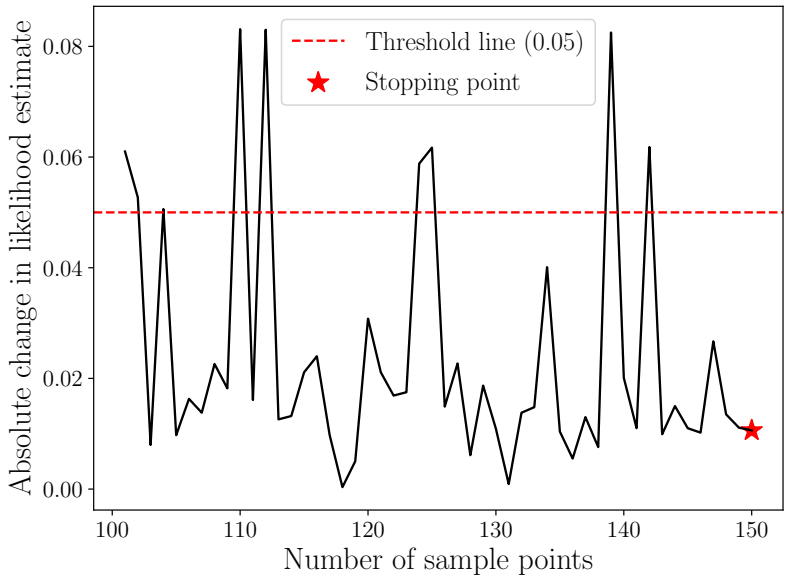
Method	N_s^+	L_T (s)	R^2	N_s^+	L_T (s)	R^2
		Branin			Himmelblau	
AL_{MC}	13	17.99	0.9999	11	21.87	0.9998
$AL_{\alpha=0.01}$	11	102.15	0.9999	21	181.58	0.9999
$AL_{\alpha=0.05}$	10	94.47	0.9994	2	22.33	0.8219
		Camel back			Ackley	
AL_{MC}	19	19.65	0.9977	128	288.05	0.9077
$AL_{\alpha=0.01}$	4	28.62	0.9154	55	15133.75	0.8429
$AL_{\alpha=0.05}$	4	27.04	0.9130	8	1420.70	0.7664

A robust and precise convergence can be reached by using multiple criteria

- With 20 initial sample points



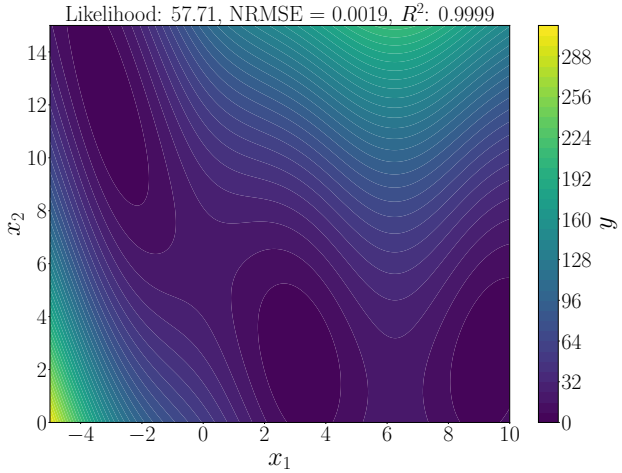
Camel back



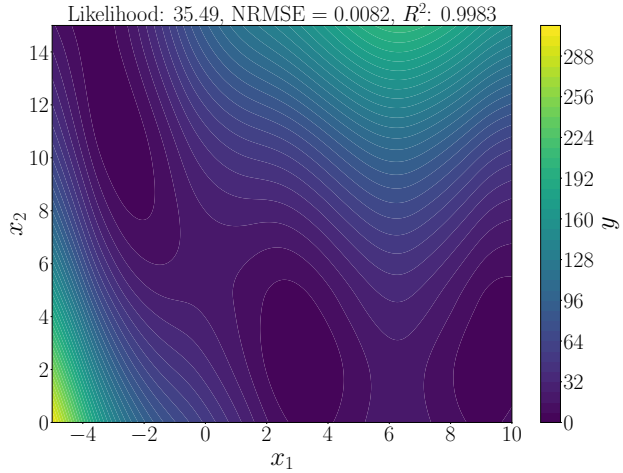
Ackley (100-150 points)

Method	N_s^+	L_T (s)	R^2	N_s^+	L_T (s)	R^2
		Branin			Himmelblau	
AL_{MC}	13	17.99	0.9999	11	21.87	0.9998
$AL_{\alpha=0.01}$	11	102.15	0.9999	21	181.58	0.9999
$AL_{\alpha=0.05}$	10	94.47	0.9994	2	22.33	0.8219
		Camel back			Ackley	
AL_{MC}	19	19.65	0.9977	128	288.05	0.9077
$AL_{\alpha=0.01}$	4	28.62	0.9154	55	15133.75	0.8429
$AL_{\alpha=0.05}$	4	27.04	0.9130	8	1420.70	0.7664

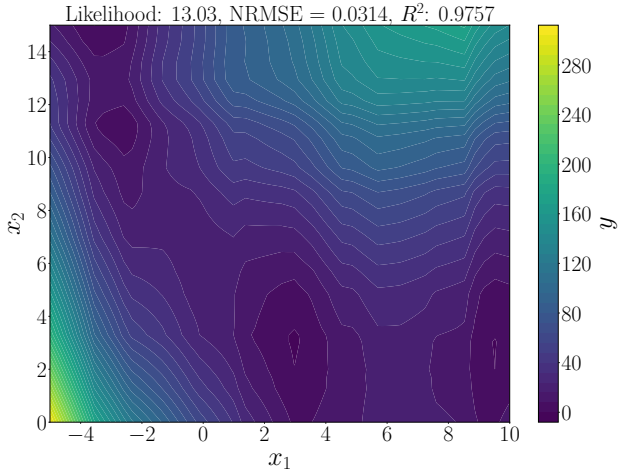
Most of the kernels have good performance on learned Branin function



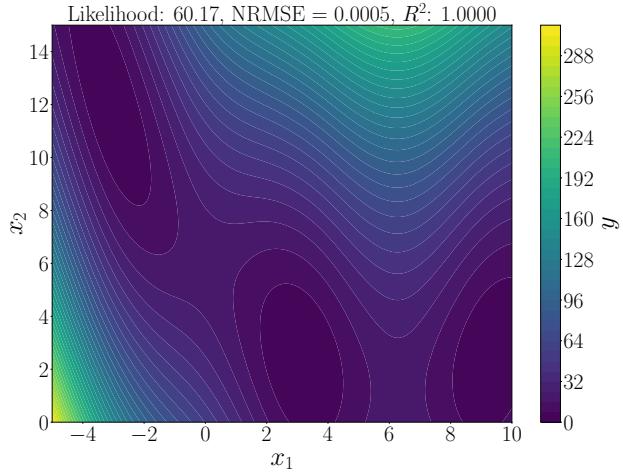
Gaussian



Matérn 3/2



Exponential



Matérn 5/2