Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/80/2024CH0075

The Application of Different Surrogate Models in Engineering Predictions

Wenqi Chen

School of Mechanical Design, Manufacturing and Automation, Dalian Maritime University, Dalian, China

24064908g@connect.polyu.hk

Abstract. Surrogate models are widely used in engineering predictions to reduce computational costs and improve efficiency, especially in complex systems where direct simulations are time-consuming and expensive. This paper explores the application of four commonly used surrogate models: Response Surface Methodology (RSM), Radial Basis Function (RBF), Kriging, and Support Vector Machine (SVM). Each model's strengths, weaknesses, and suitability for various engineering scenarios are discussed. RSM is shown to be effective for process optimization in systems with moderately nonlinear responses. RBF excels in real-time predictions and nonlinear systems, while Kriging offers high accuracy in spatial data prediction along with uncertainty quantification. SVM demonstrates strong performance in high-dimensional classification tasks. Additionally, this paper addresses strategies for reducing computational costs when applying these models, including the use of efficient optimization techniques. The findings suggest that the selection of an appropriate surrogate model depends heavily on the specific application and the complexity of the system being modeled. Future research could focus on improving the computational efficiency of these models, especially for large datasets.

Keywords: Surrogate models, engineering predictions, response surface methodology, radial basis function, kriging, support vector machine.

1. Introduction

In the process of engineering design and optimization, accurate predictions are crucial to ensuring the performance and safety of systems. However, complex engineering systems often require time-consuming and costly computations, particularly during simulations and experiments. Therefore, the development of surrogate models capable of rapidly predicting system responses has become an effective solution. By constructing surrogate models, researchers can approximate system behavior based on existing data, reducing the number of expensive computations and improving design efficiency. These models are widely used in fields such as aerospace, civil engineering, energy, and manufacturing, significantly enhancing the efficiency of design and optimization processes.

Surrogate models, also known as metamodels or proxy models, are mathematical tools used to accelerate engineering predictions by simplifying the computation of complex systems. Common surrogate models include Response Surface Methodology (RSM), Radial Basis Function (RBF), Kriging, and Support Vector Machine (SVM). These models efficiently predict system responses based on limited data and can be applied to complex multidimensional problems. Each surrogate model has its

DOI: 10.54254/2755-2721/80/2024CH0075

strengths and weaknesses in terms of performance, computational efficiency, and application scenarios. This paper discusses the application of these surrogate models in various engineering prediction contexts.

2. Types of surrogate models and their applications

2.1. RSM

RSM is widely used in optimizing industrial processes where multiple input variables influence the outcome. For example, it is commonly applied in manufacturing, chemical engineering, and pharmaceutical development to find optimal process conditions (e.g. temperature, pressure, concentration) that maximize or minimize a desired output (yield, quality, cost). RSM is valuable in situations where experimental trials are expensive or time-consuming. It allows researchers to design efficient experiments by systematically varying inputs and observing their effect on the output. This is especially useful in quality improvement initiatives or product design, where understanding the interaction between multiple factors is key. It is used in mechanical and civil engineering for creating models that predict system behavior (e.g., material strength, structural response) based on input variables. RSM helps in approximating complex functions where a precise relationship between input and output is unknown but can be modeled empirically.

RSM is particularly useful when the relationship between inputs and the response variable is nonlinear. By fitting a polynomial equation (often quadratic), RSM captures the curvature in the data to describe the response surface effectively. RSM is designed to make efficient use of experimental data. It provides a systematic approach for exploring the effects of multiple factors and their interactions without needing a large number of experiments. One of RSM's strengths is in optimization finding the set of input values that yield the best output. It can also perform sensitivity analysis, identifying which factors most significantly impact the response variable. RSM often involves graphical methods, such as contour plots or response surfaces, to visualize the relationship between input variables and the response. This makes it easier for engineers and researchers to interpret and communicate the findings. RSM assumes that the underlying relationships can be approximated by low-degree polynomials, typically quadratic. While this works well for moderate levels of complexity, it may struggle with highly complex or chaotic systems, leading to inaccuracies if the model is not properly fitted.

2.2. RBF

The RBF model has a wide range of applications in optimizing real-time monitoring and prediction in the ship design process. The RBF model for ship design optimization facilitates the real-time monitoring and prediction work as the ice resistance values are estimated very fast under different configurations of design variables. This work provides support and aids for ship design prediction with a view to improving design efficiency while ensuring improved design accuracy. The computational performance has been greatly improved, increasing the efficiency of the design process, shortening the design cycle and reducing design costs. During the real-time monitoring and prediction process, the RBF model plays a key role in improving navigation safety by instantly predicting the ice resistance as the ship passes through the ice area. The RBF model can be used to predict changes in ice resistance based on continuous monitoring of the ship's navigation status, combined with a number of parameters of the surrounding environment, to provide the ship's captain with information to formulate a better navigation strategy.

Unlike global methods such as polynomial regression, RBF has a localized influence, meaning each basis function only significantly affects predictions in a certain region around its center. This makes RBF highly adaptable to local variations in data, capturing fine-grained patterns in complex systems. One of the key strengths of RBF is its ability to generate smooth, continuous surfaces from discrete data points, making it ideal for applications like terrain modeling, surface reconstruction, and fluid dynamics.

2.3. Kriging

Kriging was originally developed for geostatistics and is widely used for interpolating and predicting values in mining, oil and gas exploration, and environmental monitoring. It can predict unknown values,

DOI: 10.54254/2755-2721/80/2024CH0075

such as mineral concentrations or pollutant levels, based on spatially distributed data points. Kriging is often used as a surrogate model (or metamodel) in aerodynamics, structural engineering, and fluid dynamics. It helps approximate expensive-to-evaluate simulations by predicting responses like pressure distributions, temperature fields, or stresses across different designs, significantly reducing computation time in optimization problems. Kriging is ideal for scenarios requiring both prediction and uncertainty quantification. It is commonly used in risk analysis and safety assessments, where it's crucial not only to predict outcomes but also to understand the confidence or uncertainty in those predictions. For example, in climate modeling or environmental impact assessments, Kriging provides a means of evaluating the reliability of predictions.

Kriging is one of the most accurate methods for spatial interpolation because it incorporates both the distance between known data points and the overall spatial correlation. This allows it to make smooth and realistic predictions that honor the observed data while minimizing error. A distinctive feature of Kriging is that it provides not only a prediction but also an estimation of the uncertainty (variance) associated with that prediction. This makes it especially valuable in fields where understanding the confidence in predictions is critical, such as resource estimation in mining or environmental risk assessment. Although highly accurate, Kriging can be computationally intensive, especially when working with large datasets [1]. The complexity arises from the need to solve a system of equations based on the covariance matrix, which grows with the number of data points.

2.4. SVM

SVM is widely used in binary and multi-class classification tasks, making it ideal for applications in image recognition, text categorization, bioinformatics, and handwriting recognition. It excels in distinguishing between classes by finding the optimal hyperplane that separates data points. SVM can also be applied to regression problems, where it predicts continuous values. SVR is used in fields like financial forecasting, energy load prediction, and environmental modeling to predict values based on complex datasets. SVM is effective in detecting outliers or anomalies, making it suitable for applications like fraud detection, network intrusion detection, and fault detection in engineering systems.

One of the major strengths of SVM is its ability to handle high-dimensional data efficiently [2]. This makes it ideal for tasks where the number of features is large compared to the number of data points, such as in genomics or text classification. SVM tends to perform well in cases where the number of dimensions exceeds the number of samples, as it relies on maximizing the margin between classes. This characteristic helps SVM generalize well to new data, particularly in smaller datasets with complex boundaries. SVM works best when there is a clear margin of separation between classes. In cases where the data is highly overlapping or noisy, other models like Random Forests or Neural Networks may perform better.

3. Comparison of surrogate models

3.1. Models

3.1.1. RSM. RSM is a statistical technique used to explore the relationships between one or more response variables (outputs) and several input variables (factors). It is typically used for optimizing processes by fitting a regression model, usually a second-order polynomial model, to the data. The general formula for RSM is as below Equation 1:

$$y = \beta_0 + \sum_{i-1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon \quad (1)$$

RSM offers the advantage of efficiently optimizing multi-factor systems by reducing experimental runs, providing visual insights through response surface plots, and focusing on local optimization. However, its reliance on quadratic models limits it to local regions, and the model complexity increases with more factors, potentially leading to overfitting. Additionally, RSM's assumptions about error normality and model continuity may not suit all systems, particularly highly nonlinear or global ones.

DOI: 10.54254/2755-2721/80/2024CH0075

3.1.2. RBF. RBF is a real-valued function whose value depends only on the distance from a central point, often used in machine learning and interpolation. RBFs are commonly employed in neural networks, support vector machines (SVM), and function approximation. The general form of an RBF is shown as:

$$\Phi(\| x - c \|) \tag{2}$$

RBFs are highly flexible for function approximation and interpolation, offering the advantage of universal approximation and local influence, making them effective for smooth, localized modeling. They are also easy to implement in neural networks, often requiring fewer training examples. However, RBFs can become computationally expensive as the dataset grows, and their performance is sensitive to parameter choices like the spread in Gaussian RBFs, which can be challenging to optimize. Additionally, they may overfit noisy data if not properly regularized, making parameter tuning crucial for effective use.

3.1.3. Kriging. Kriging is a geostatistical interpolation method used to predict values of a spatially distributed variable at unmeasured locations based on observed data points. It is widely applied in fields such as geostatistics, environmental science, mining, and meteorology for tasks like mapping and spatial prediction. Kriging not only estimates the variable at unknown points but also provides an estimate of the error or uncertainty associated with each prediction.

The general Kriging model is shown as Equation 3:

$$Z(s) = \mu(s) + \varepsilon(s) \tag{3}$$

Kriging offers the advantage of providing the best linear unbiased predictions by accounting for spatial autocorrelation, making it more accurate than simpler interpolation methods. It also estimates prediction uncertainty, which is valuable for decision-making and risk assessment. However, Kriging is computationally intensive, especially with large datasets, and its accuracy heavily depends on correctly estimating the variogram. It can also be sensitive to outliers and requires expertise to apply and interpret, making it more complex than other spatial interpolation methods.

3.1.4. SVM. SVM is a supervised machine learning algorithm primarily used for classification tasks but can also be applied to regression problems. It works by finding a hyperplane that best separates data points into different classes in a high-dimensional space. SVM aims to maximize the margin between the hyperplane and the nearest data points from each class, which are called support vectors. The larger the margin, the better the generalization ability of the model.

For linearly separable data, SVM solves the following optimization problem as shown in Equation 4:

$$\text{Minimize } \frac{1}{2} \| \mathbf{w} \|^2 \quad \text{subject to} y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i \tag{4}$$

SVM excel in handling high-dimensional data and creating clear decision boundaries by maximizing the margin, reducing the risk of overfitting. They are highly flexible due to the use of kernel functions, which allow for the effective handling of non-linear data according to Table 1. However, SVMs can be computationally expensive, especially with large datasets, and require careful tuning of parameters like the regularization term and kernel settings for optimal performance. Additionally, SVM models, particularly with non-linear kernels, can be difficult to interpret.

DOI: 10.54254/2755-2721/80/2024CH0075

Table 1. Comparison of Surrogate Models.

Model	Accuracy	Computational Efficiency	Best Engineering Scenarios
RSM	Moderate	High	Process optimization in chemical and manufacturing processes
RBF	High for nonlinear systems	Moderate (Low with large data)	Real-time monitoring and optimization tasks (e.g., ship design)
Kriging	Very High (spatial data)	Low (large datasets)	Geostatistics, resource estimation, aerodynamic design
SVM	High (complex/high-dim data)	Moderate to Low (large datasets)	Classification, fault detection, image recognition, medical diagnosis

3.2. Scenario-specific applications

Recent advancements in the application of Artificial Neural Network (ANN)-based proxy models have shown remarkable success in various engineering prediction scenarios. Musayev, K. applied an ANN-based proxy model to optimize well placement for pressure management in Geological Carbon Storage (GCS). The model's effectiveness was demonstrated by optimizing the locations of CO2 injection and brine extraction wells in the Pohang Basin's upper aquifer, where the best-performing ANN model was selected through rigorous testing. This model, combined with a genetic algorithm, identified the optimal well locations to maximize cumulative CO2 injection [3]. Similarly, Aydin, H. developed a proxy model that estimates reservoir pressure and temperature using wellhead data such as pressure, temperature, and non-condensable gas (NCG) levels. This model was trained with a comprehensive dataset generated from a calibrated wellbore simulator [4]. In another study, Panjalizadeh, H. introduced a dynamic ANN-based proxy model to optimize steam injection processes in heavy oil reservoirs. By coupling a transient ANN with a genetic algorithm, the study successfully identified the optimal steam injection rate, steam quality, and injection timing. These studies underscore the versatility and effectiveness of ANN-based proxy models in addressing complex engineering challenges, providing a robust framework for optimizing various processes [5].

3.3. Optimization of surrogate models

3.3.1. Parameter tuning. Optimizing surrogate models from the perspective of parameter tuning can be summarized in Table 2.

Table 2. Surrogate models’ optimization.

Model	Parameter Types	Optimization Techniques	Common Applications
RBF	Kernel parameters, Regularization terms	Grid Search, Random Search, Gradient-Based Optimization	Real-time prediction, non-linear regression
RSM	Factors in experimental design, Response surface parameters	Central Composite Design, Steepest Descent, Newton’s Method	Process optimization, experimental design
Kriging	Covariance function parameters, Noise variance	Maximum Likelihood Estimation, Bayesian Optimization	Spatial data modeling, geostatistics
SVM	C Parameter, Kernel parameters (e.g., gamma)	Grid Search, Random Search, Cross-Validation	Classification, pattern recognition, regression

3.3.2. Reducing computational costs. Reducing computational costs is essential for efficient model optimization, especially when dealing with complex models or large datasets [6]. To achieve this, several strategies can be employed. First, choosing efficient algorithms, such as Stochastic Gradient Descent (SGD) instead of full-batch gradient descent, can significantly cut down on computation time by processing smaller data subsets [7]. Simplifying models through feature selection, model pruning, or opting for simpler models can also reduce complexity and computational demands. Efficient data handling techniques, including data sampling and compression, help manage large datasets effectively. Additionally, leveraging parallel and distributed computing resources, such as multi-core processors, GPUs, or cloud-based solutions, can accelerate computations. For model optimization, techniques like Bayesian optimization for hyperparameter tuning and early stopping during training can further minimize computational costs [8]. Implementing caching to store intermediate results and reusing pre-trained models or applying transfer learning can also contribute to cost reduction. Specifically for RBF models, approximate kernel evaluations and sparse approximations can improve efficiency, while RSM can benefit from efficient experimental design techniques [9]. In Kriging models, low-rank approximations and sparse Gaussian processes help handle large-scale data, and for SVM, approximate methods and incremental learning techniques can reduce the computational burden [10]. By combining these approaches, significant reductions in computational costs can be achieved, enhancing overall efficiency and performance.

4. Conclusion

This paper analyzed and compared four commonly used surrogate models: RSM, RBF, Kriging, and SVM. Each model exhibits specific strengths based on the engineering application. RSM is highly efficient for optimizing processes with smooth, low-order polynomial relationships, making it suitable for smaller, well-behaved systems. RBF demonstrates strong performance in capturing nonlinearities and local variations, which is advantageous for real-time prediction tasks, although its efficiency decreases with larger datasets. Kriging stands out for its high accuracy in predicting spatially correlated data and its ability to quantify uncertainty, but its computational demands limit its use with larger datasets. SVM is effective in high-dimensional classification and regression tasks, particularly when clear margins exist between classes, though its computational cost becomes a challenge with large datasets. Overall, the selection of the appropriate surrogate model depends on the system's complexity and the specific requirements of the engineering problem.

Several future research directions could further enhance the performance and applicability of surrogate models in engineering predictions. Improving computational efficiency is a primary focus, especially for handling large datasets and complex systems. Techniques such as sparse approximations for Kriging and more efficient kernel evaluations for RBF could help address this challenge. Another promising area is the integration of emerging machine learning techniques, such as deep learning, which could enhance the predictive power and generalization capabilities of surrogate models. Additionally, developing adaptive surrogate models that can adjust dynamically to new data or changing system conditions is crucial for applications involving real-time monitoring and evolving environments. Multifidelity approaches that combine high-fidelity, high-cost simulations with lower-fidelity, lower-cost models could also strike a balance between accuracy and computational efficiency. By pursuing these directions, the potential for surrogate models to solve increasingly complex engineering problems will be significantly expanded.

References

[1] Musayev, K., Shin, H., & Nguyen-Le, V. (2023). Optimization of CO2 injection and brine production well placement using a genetic algorithm and artificial neural network-based proxy model. International Journal of Greenhouse Gas Control, 127, 103915.

[2] Aydin, H., Akin, S., & Senturk, E. (2020). A proxy model for determining reservoir pressure and temperature for geothermal wells. Geothermics, 88, 101916.

DOI: 10.54254/2755-2721/80/2024CH0075

[3] Panjalizadeh, H., Alizadeh, N., & Mashhadi, H. (2014). A workflow for risk analysis and optimization of steam flooding scenario using static and dynamic proxy models. Journal of Petroleum Science and Engineering, 121, 78-86.

[4] Di Lullo, G., Oni, A. O., & Kumar, A. (2023). The development of complex engineering models using artificial neural network-based proxy models for life cycle assessments of energy systems. Renewable and Sustainable Energy Reviews, 184, 113583.

[5] Zhao, Y., Lu, H., Fang, K., Zhang, P., Chen, D., Bräuning, A., Grießinger, J., Yang, B., Liang, C., Zhang, H., Sun, Y., Jin, Q., Li, J., & Liu, X. (2024). Deciphering the transfer of hydroclimate signals to tree-ring δ18O using a proxy system model in East Asia's Meiyu region. Chemical Geology, 668, 122350.

[6] Zhang, F., Nghiem, L., & Chen, Z. (2023). Evaluating reservoir performance using a transformer based proxy model. Geoenergy Science and Engineering, 226, 211644.

[7] Kim, J., Lee, K., & Choe, J. (2021). Efficient and robust optimization for well patterns using a PSO algorithm with a CNN-based proxy model. Journal of Petroleum Science and Engineering, 207, 109088.

[8] Peng, X., Rao, X., Zhao, H., Xu, Y., Zhong, X., Zhan, W., & Huang, L. (2022). A proxy model to predict reservoir dynamic pressure profile of fracture network based on deep convolutional generative adversarial networks (DCGAN). Journal of Petroleum Science and Engineering, 208(Part C), 109577.

[9] Rigo, C. A., Seman, L. O., Morsch Filho, E., Camponogara, E., & Bezerra, E. A. (2023). MPPT aware task scheduling for nanosatellites using MIP-based ReLU proxy models. Expert Systems with Applications, 234, 121022.

[10] Zhong, Z., Sun, A. Y., Wang, Y., & Ren, B. (2020). Predicting field production rates for waterflooding using a machine learning-based proxy model. Journal of Petroleum Science and Engineering, 194, 107574.