Multivariate Active Learning for Polynomial Chaos Expansion

Qitian Lu, Miroslav Vořechovský, Lukáš Novák
Faculty of Civil Engineering, Brno University of Technology, Brno 602 00, Czechia,
Lukas.Novak4@vut.cz

Abstract. The paper extends the active learning strategy of (Novák et al., 2021), which was originally designed for scalar-output polynomial chaos expansion surrogate models, to the case of vector-valued quantities of interest. Physical systems simulated by mathematical models typically contain multiple outputs (e.g., extrema in stress fields), and thus it is often necessary to construct several surrogate models for subsequent uncertainty quantification of each output. While the construction of surrogate models one-by-one for a single one-shot experimental design generated by various techniques (e.g., Latin Hypercube Sampling) is straightforward, an active learning methodology for multiple outputs remains an open research question. The proposed method combines the adaptivity of a polynomial chaos expansion with sequential sampling, enabling a one-by-one extension of the experimental design. The iterative process of sequential sampling selects points from a large pool of candidates by attempting to cover the design domain proportionally to their local variance contribution. The proposed criterion for sample selection balances between exploitation of the surrogate model and exploration of the design domain, while accounting for the different contributions of each output to the total criterion. The obtained numerical results confirm its superiority over standard non-sequential approaches in terms of surrogate model accuracy and estimation of output variance, especially with respect to extreme values.

Keywords: Uncertainty Quantification, Active Learning, Polynomial Chaos Expansion

1. Motivation

Uncertainty quantification of the quantity of interest (QoI), the outcome of a physical system represented by a complex mathematical model, is typically computationally costly; thus, it is often necessary to construct a surrogate model. One of the most suitable surrogate models for UQ is polynomial chaos expansion (PCE), which allows for analytical estimation of statistical moments and sensitivity analysis (Novák, 2022; Sudret, 2008). However, the quality of the approximation is governed by the quality of the experimental design (ED) used for the construction of non-intrusive PCE. On the one hand, it would be optimal to use a very large set of samples for ED, but on the other hand, each sample brings an additional significant computational burden (e.g., nonlinear finite element method); thus, it is necessary to reduce the number of model evaluations as much as possible in the process of training the surrogate model while maintaining the accuracy of the approximation. This challenge is commonly addressed by advanced sampling strategies (Shields et al., 2015; Fajraoui et al., 2017; Hadigol and Doostan, 2018). Moreover, active learning approaches provide a more targeted strategy, iteratively enriching the experimental design based on prescribed criteria (Jones et al., 1998; Marelli and Sudret, 2018). Recently, Novák et al. (Novák et al., 2021) proposed Θ criterion based on simple idea of combining exploration aspect with variance density obtained directly from PCE, thereby achieving a balance between space-filling and variance-based model refinement. This strategy has been demonstrated to improve surrogate accuracy and provide more reliable estimation of the output variance compared to standard non-sequential sampling approaches while introducing only negligible additional computational cost.

Despite these merits, the original Θ criterion is formulated for single-output problems. This restriction may be limiting in many engineering applications, where one simulation simultaneously produces several outputs and all of them must be approximated under a limited and shared computational budget. The straightforward single-output criterion is generally unsatisfactory for several reasons. First, using different ED for each of the outputs may make it difficult to maintain consistency across the surrogate construction, which is undesirable when multiple responses are to be approximated jointly in a unified multi-output setting. Second, the repeated sampling procedure for different training sets may become computationally inefficient when the underlying model is expensive. Third, different outputs may exhibit different levels of variability, smoothness, and approximation difficulty, so that a sample selected as informative for one output may be much less informative for other outputs. Therefore, the extension from single-output to multi-output criterion is not merely a formal generalization but a necessary step toward an efficient sampling strategy for practical multi-output problems.

Motivated by this limitation, this paper proposes a multi-output Θ criterion for the adaptive sequential sampling for non-intrusive PCE. The proposed approach aggregates variance and distance information from multiple outputs into a unified criterion, so that each additional sample is selected to maximize the overall improvement of the surrogate model. The proposed method is verified through several numerical examples. The results show that the proposed method provides an effective and efficient sampling strategy for multi-output PCE surrogate models.

2. Non-intrusive Polynomial Chaos Expansion

Let $Y = g(\boldsymbol{X})$ represent a computational model, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_M)$ is a set of M independent random inputs. Under the Wiener-Askey framework (Xiu and Karniadakis, 2002), the input variables are transformed into a series of standardized germs $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_M)$, with respect to which an orthonormal polynomial basis can be constructed. The model response is then approximated by a truncated expansion in terms of these basis functions, ensuring a finite representation suitable for practical computation, as

$$Y \approx \sum_{\alpha \in \mathcal{A}} \beta_\alpha \Psi_\alpha(\boldsymbol{\xi})$$

where $\alpha \in \mathbb{N}^M$ denotes a multi-index specifying the polynomial degree associated with each germ dimension ξ_i, $\mathcal{A}$ is the truncated multi-index set, β_α are the corresponding expansion coefficients, and $\Psi_\alpha(\boldsymbol{\xi})$ are multivariate orthonormal polynomial bases constructed from a tensor product of univariate orthonormal polynomials

$$\Psi_\alpha(\boldsymbol{\xi}) = \prod_{i=1}^{M} \psi_{\alpha_i}(\xi_i)\,.$$

The coefficients β_α can be estimated using regression techniques (Blatman and Sudret, 2011). Owing to the orthonormality of the basis functions, the variance of the response σ_Y^2 can be directly obtained, given by

$$\sigma_Y^2 = \sum_{\substack{\alpha\in\mathcal{A}\\ \alpha\neq 0}} \beta_\alpha^2.$$

The variance can be regarded as an integral of local contributions over the input space. Based on the PCE representation, a local variance measure can therefore be defined to quantify the contribution of a given point ξ to the total variance. In particular, the local variance is expressed as

$$\sigma_{\mathcal{A}}^2(\boldsymbol{\xi}) = \left[\sum_{\substack{\alpha\in\mathcal{A}\\ \alpha\neq 0}} \beta_\alpha\,\Psi_\alpha(\boldsymbol{\xi})\right]^2 p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$$

where $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ represents the probability density function of the standardized germs. A large local variance indicates that the surrogate carries greater uncertainty in the corresponding region, and is therefore of higher importance for sampling.

3. Adaptive Sequential Sampling

3.1. Extension of Variance-based Sequential Sampling

Once the PCE surrogate is constructed, it can be exploited to guide the selection of new samples in an adaptive manner. In particular, the experimental design is iteratively enriched by selecting new points according to a variance-based criterion derived from the surrogate model. For the single-output case, the Θ criterion is defined as (Novák et al., 2021)

$$\Theta(\boldsymbol{\xi}^{(c)}) \equiv \Theta_c = \underbrace{\sqrt{\sigma_c^2 \cdot \sigma_s^2}}_{\text{ave. var. density}}\ \underbrace{l_{c,s}^M}_{\text{vol.}}$$

where σ_c^2 and σ_s^2 are the local variance $\sigma_{\mathcal{A}}$ evaluated at the candidate point $\boldsymbol{\xi}^{(c)}$ and its nearest neighbor $\boldsymbol{\xi}^{(s)}$, respectively, and $l_{c,s}$ is the distance between these two points in the input space. The first term reflects the local contribution to the response variability, while the second term indicates spatial exploration of the design domain. The combination of these two terms ensures a balance between refining regions of high variability and maintaining a proper coverage of the input space.

In the multi-output setting, the local variance must be extended to account for multiple quantities of interest, as they may differ significantly in scale, leading to disproportionate contributions to the overall variability. For this purpose, a unified variance measure is constructed by aggregating the contributions from all outputs. The general form of multi-output Θ criterion is formulated as

$$\Theta_c = \underbrace{\sqrt{\left(\sum_{i=1}^{N_\mathrm{e}} w_{c,i}\lambda_{c,i}\right) \cdot \left(\sum_{i=1}^{N_\mathrm{e}} w_{s,i}\lambda_{s,i}\right)}}_{\text{ave. var. density}} \underbrace{l_{c,s}^M}_{\text{vol.}}$$

where $w_{c,i}$ and $w_{s,i}$ are weight factors of the ith eigenvalue (PCE function) for the candidate point and the nearest selected point, respectively, introduced to prevent dominance of large-scale outputs and other complicating factors. To obtain equal weights for individual eigenmodes (PCEs), one can select $w_{c,i} = \max_D \lambda_{c,i}$ and $w_{s,i} = \max_D \lambda_{s,i}$. Moreover, in case the individual models (PCEs) are independent (uncorrelated), the eigenvalue $\lambda_{c,i}$ are equal to the local variance density $\sigma_{c,i}^2$ of PCE_i at the location of candidate c. In this case, the multi-output Θ criterion can be written as

$$\Theta_c = \sqrt{\left(\sum_{i=1}^{N_\mathrm{PCE}} \frac{\sigma_{c,i}^2}{\max_D \sigma_{c,i}^2}\right) \cdot \left(\sum_{i=1}^{N_\mathrm{PCE}} \frac{\sigma_{s,i}^2}{\max_D \sigma_{s,i}^2}\right)}\, l_{c,s}^M.$$

In this work, no truncation is applied, and N_e equals N_PCE. It is worth noting that, when all eigenvalues are equally weighted (i.e., $w = 1$) and outputs are correlated, using only the sum of the eigenvalues of the covariance matrix leads to a result equivalent to that obtained under the assumption of independent outputs. This is because the sum of eigenvalues equals the trace of the covariance matrix, which corresponds to the sum of marginal variances and does not capture cross-correlations between outputs. Therefore, to account for correlations among outputs, additional weighting strategies are required. In this context, the correlation matrix provides a natural reference for constructing such weights, enabling the aggregation to reflect the dependency relationship between different quantities of interest. However, in our practice, although the coupling may influence the specific selection of samples during the sequential process, the overall advantage of the Θ-criterion over non-sequential sampling strategies generally remains robust. As the number of added samples increases, the selected sample set tends to progressively approach a similar global distribution.

4. Numerical Experiments

4.1. Illustrative 2D Example

To illustrate the effectiveness of the proposed multi-output Θ criterion, we consider a 2D line-singularity function defined over the unit square $(X_1, X_2) \in [0,1]^2$, which exhibits point symmetry with respect to the center $(0.5, 0.5)$. The model is given by

$$f(X_1, X_2) = \frac{1}{|0.3 - X_1^2 - X_2^2| + \delta} - \frac{1}{|0.3 - (1-X_1)^2 - (1-X_2)^2| + \delta}, \qquad (X_1, X_2) \in [0,1]^2.$$

The singular behavior occurs along circular curves where the denominators of two terms approach equal to δ, a regularization parameter to control the sharpness of the singular regions, leading to sharply localized gradients confined to narrow regions. Such localized features pose a challenge for PCE, particularly when the parameter δ is small. In this study, we consider $\delta = 0.1$, for which the response variance is approximately $\sigma_Y^2 \approx 13.070\,477$. Due to the concentration of variability in limited regions, conventional space-filling designs may fail to allocate sufficient samples near the singular peak, resulting in reduced surrogate accuracy. Moreover, to transform the problem into a multi-output case, we introduce a partition of the input domain along the diagonal (see Figure 1), thereby defining two complementary mathematical models:

$$Y_1(X_1, X_2) = \begin{cases} f(X_1, X_2), & X_1 \geq X_2, \\ 0, & X_1 < X_2, \end{cases} \qquad Y_2(X_1, X_2) = \begin{cases} 0, & X_1 > X_2, \\ f(X_1, X_2), & X_1 \leq X_2. \end{cases}$$

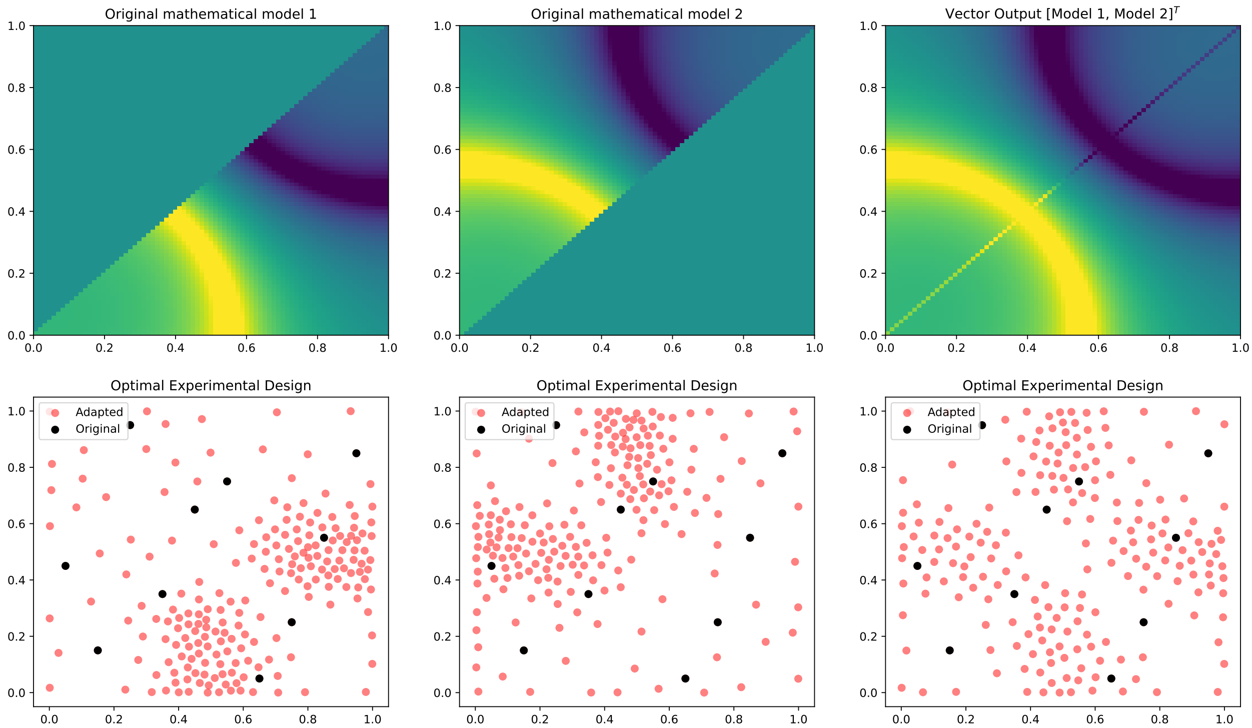

Figure 1. 2D line-singularity function and corresponding multi-output formulation with optimal experimental designs.

As a baseline sampling strategy, Latin Hypercube Sampling (LHS) is considered due to its widespread use in uncertainty quantification (Choi et al., 2004). Figure 2 summarizes the performance over 100 independent runs in terms of the mean and standard deviation of the log-transformed maximum absolute error (MaxAE), mean absolute error (MAE), relative absolute error of variance, and leave-one-out (LOO) error. As illustrated in Figure 2(a), the Θ criterion achieves a noticeably faster reduction in MaxAE at early stages and maintains lower error throughout the

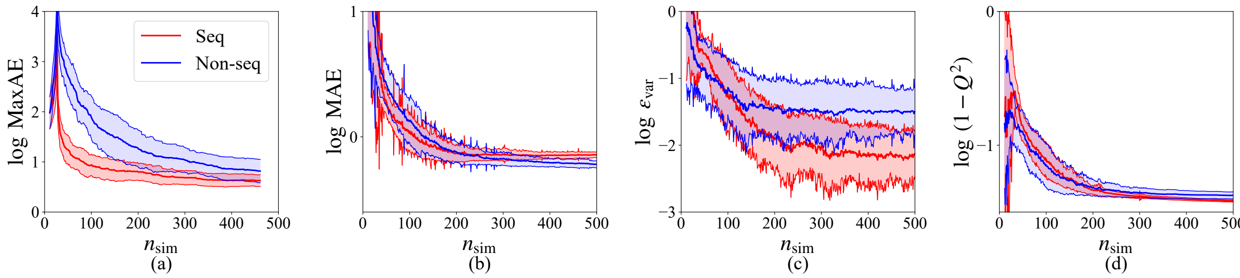

Figure 2. 2D line-singularity function: comparison of Θ criterion and LHS for PCE surrogate model construction: mean and standard deviation of (a) MaxAE, (b) MAE, (c) relative variance error, and (d) LOO error.

whole sampling process compared to the LHS. Although the gap gradually becomes narrow as the sample size increases, the Θ criterion consistently exhibits a tighter standard deviation region, indicating better robustness in the worst-case prediction. For MAE shown in Figure 2(b), a similar trend can be observed at the early stage. When n_{sim} reaches around 250, the performance of two strategies becomes comparable. After this point, the Θ-based method exhibits a slight increase in MAE, however, it still retains a narrower spread, reflecting more stable performance across repeats. In terms of the relative variance error in Figure 2(c), the advantage of the Θ-based strategy becomes more pronounced. The variance error decreases faster and converges to a consistently lower level than that of non-sequential LHS. This behavior suggests that the multi-output Θ criterion accounts for the local variance contributions from both outputs, which effectively improves the estimation of second-order statistics in the surrogate for the complete model. For the LOO error in Figure 2(d), both methods show a similar overall decreasing trend, but finally the Θ criterion achieves a relatively lower error and significantly lower variability.

4.2. Metaball function – complicated topology

The second example is a metaball function (Vořechovský, 2022), which consists of a superposition of two smooth components and can be naturally reformulated as a multi-output problem by treating the two components as separate outputs. The interaction between these components produces a highly nonlinear and non-convex failure surface, with multiple regions of important influence, making it challenging for surrogate modeling and adaptive sampling, given by

$$g\left(x_1, x_2\right)=\frac{30}{\left(\frac{4\left(x_1+2\right)^2}{9}+\frac{x_2^2}{25}\right)^2+1}+\frac{20}{\left(\frac{\left(x_1-2.5\right)^2}{4}+\frac{\left(x_2-0.5\right)^2}{25}\right)^2+1}-5. \tag{1}$$

As shown in Fig. 3, the sequential design based on the multi-output Θ criterion shows a better performance than the non-sequential approach across both components and the complete model. From the perspective of surrogate accuracy, the Θ criterion effectively controls extreme errors while maintaining comparable levels of MAE and relative variance error for both individual components and the complete model. In particular, both MaxAE and MAE exhibit narrower standard deviation regions in the later stages, indicating that the samples selected by the Θ criterion are less sensitive to

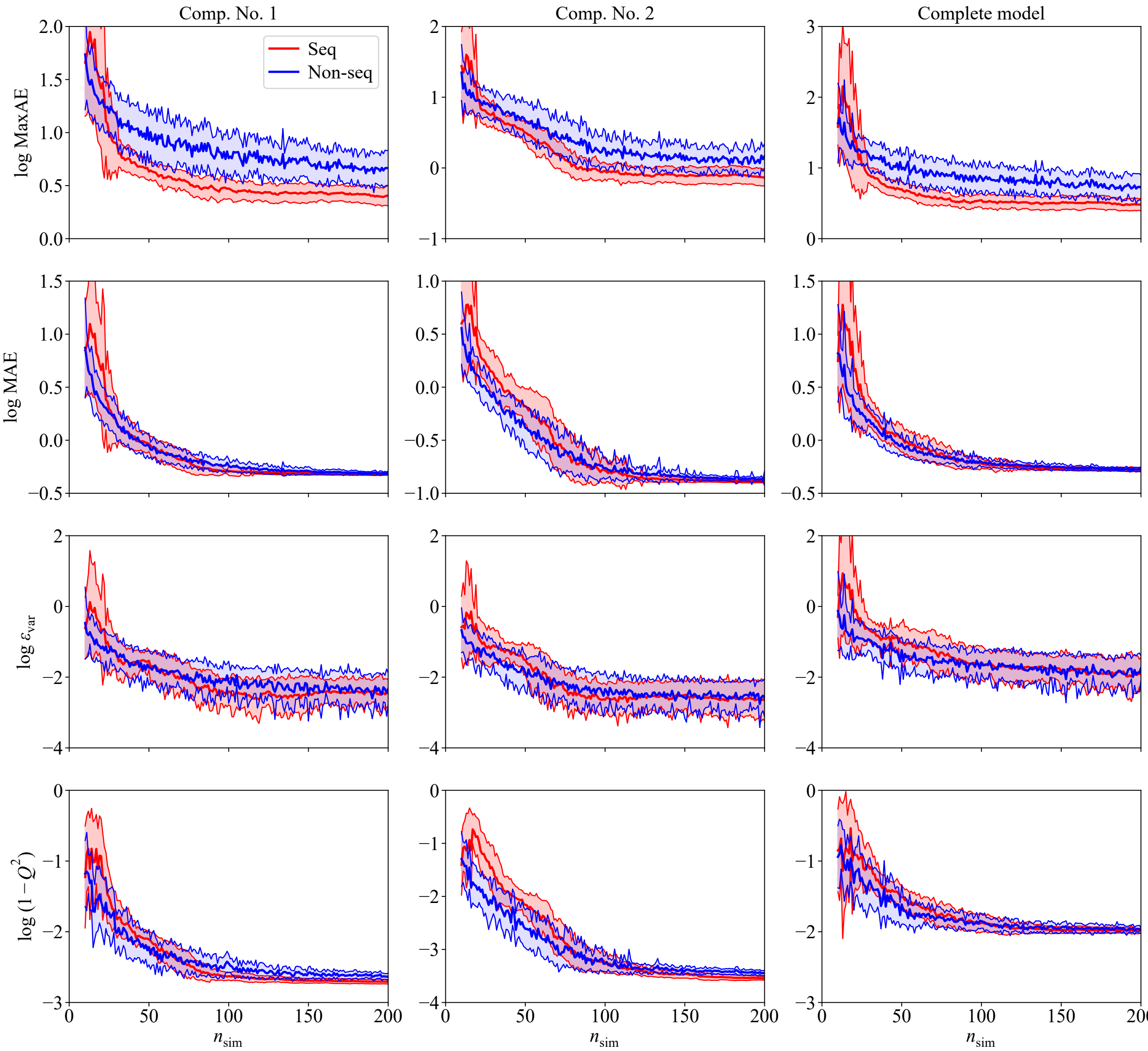

Figure 3. Metaball function: comparison of Θ criterion and LHS for PCE surrogate model construction: mean and standard deviation of MaxAE, MAE, relative variance error, and LOO error

randomness in repeated runs. From the perspective of surrogate stability, the model using samples selected by the Θ criterion shows a slight advantage for the individual components, with lower mean LOO errors and narrower standard deviation bands. For the full model, the Θ criterion maintains a level of stability comparable to that of the non-sequential LHS.

4.3. Four Branches problem

The last example in this paper is a 2D Four Branches function (Borri and Speranzini, 1997), which is a common benchmark problem in reliability analyses; see, e.g., studies with various parameter settings (Schueremans and Gemert, 2005; Echard et al., 2011; Papaioannou et al., 2016; Schöbi et al., 2017). The function describes the failure of a series system, expressed as

$$g_1\left(x_1, x_2\right) = \min \begin{cases} 3 + 0.1\left(x_1 - x_2\right)^2 - \left(x_1 + x_2\right)/\sqrt{2} \\ 3 + 0.1\left(x_1 - x_2\right)^2 + \left(x_1 + x_2\right)/\sqrt{2} \end{cases} \tag{2}$$

$$g_2\left(x_1, x_2\right) = \min \begin{cases} x_1 - x_2 + 7/\sqrt{2} \\ x_2 - x_1 + 7/\sqrt{2} \end{cases} . \tag{3}$$

In this formulation, the system can be interpreted as consisting of two failure modes, represented by g_1 and g_2, corresponding to nonlinear and linear branches of the failure surface, respectively. To investigate the performance of the proposed multi-output Θ criterion, two separate PCE surrogate models are constructed for g_1 and g_2 while sharing the same experimental design. The combined performance function is defined as $g = \min\{g_1, g_2\}$, and failure occurs when $g(x_1, x_2) \leq 0$. Due to the symmetry of the problem, both outputs exhibit similar behavior.

This case is more challenging for PCE than the previous smooth examples, because both component functions and the complete model are defined through minimum operators. These operators introduce non-smooth switching regions, where the active branch changes. Such piecewise behavior cannot be represented efficiently by a global polynomial basis, and therefore may lead to slower and less stable convergence of the PCE approximation. Despite this difficulty, the sequential design based on the multi-output Θ criterion remains effective (see Figure 4). For MaxAE, the multi-output Θ criterion shows a clear reduction for each component and the complete model, which is consistent with the previous two examples. For MAE, the two strategies show comparable overall performance. For the first component, the Θ criterion exhibits larger fluctuations in the intermediate stages, but its performance becomes comparable to LHS near the end. Notably, for the complete model, the Θ criterion shows a certain advantage over the non-sequential approach, indicating that the adaptively selected samples can still be beneficial when the component functions are combined. For the relative variance error, the trend is different. In the complete model, the Θ criterion leads to a larger relative variance error than the non-sequential LHS, although it exhibits a narrower error band. The reason for this phenomenon is that, unlike the previous two examples, where the complete models are formed by combining outputs with different input-domain characteristics, the minimum operator applied to g_1 and g_2 generates additional narrow regions with high local variance near the switching boundaries. These regions are not explicitly emphasized by the Θ criterion and may therefore be overlooked during the sequential selection process, whereas the space-filling nature of LHS makes it more likely to have a small number of samples in these regions. For the first component, the Θ criterion still performs relatively well. Regarding model stability, the results are consistent with the previous examples, i.e. the surrogate model using the samples from the Θ criterion exhibits better stability than that based on LHS.

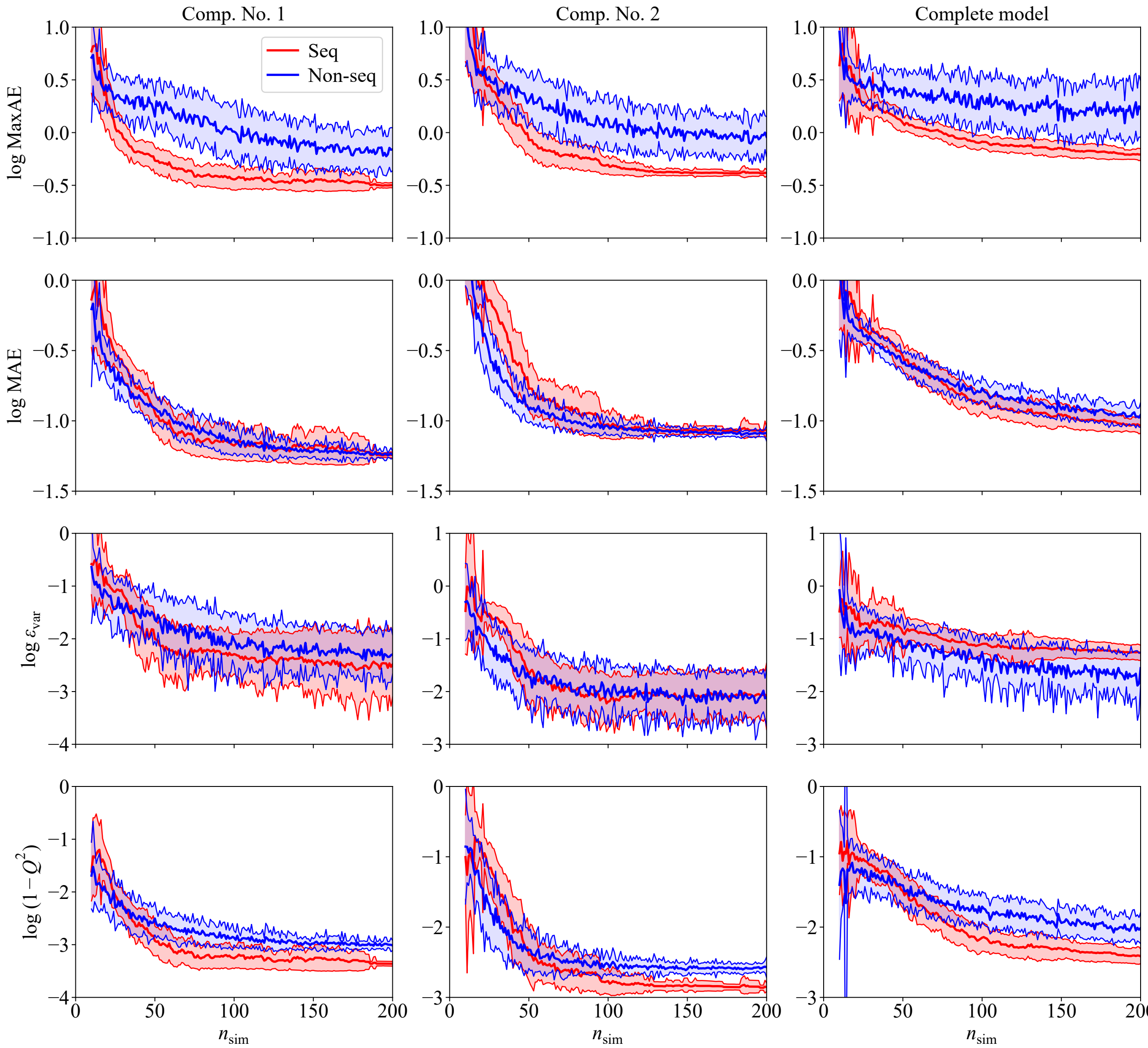

Figure 4. 2D Four Branches function: comparison of Θ criterion and LHS for PCE surrogate model construction: mean and standard deviation of MaxAE, MAE, relative variance error, and LOO error

5. Conclusion

This work extends the variance-based Θ criterion to a multi-output setting for the adaptive sampling of PCE surrogates. By aggregating the local variance contributions of multiple quantities of interest into a unified criterion, the proposed multi-output Θ criterion enables a robust and reliable sequential enrichment of a single experimental design shared across all outputs. The numerical results show that the proposed criterion consistently outperforms non-sequential sampling strategies across all examples, achieving higher accuracy and improved robustness in various metrics. In particular, it provides more reliable estimation in worst-case scenarios (measured by maximum absolute error), suggesting its potential for reducing sampling costs in multi-output engineering problems where the extreme value is of primary concern.

It should be noted that the present study assumes independence among the surrogate models associated with different outputs. In practice, however, correlations between quantities of interest are often present and may influence the effect of sampling strategy. As discussed in Section 3, the incorporation of weighting parameters provides a potential way to account for such dependencies, with the correlation matrix serving as a natural reference. Another possibility is to apply a direct decorrelation to the surrogate outputs, for instance, based on the covariance structure of surrogate prediction at each iteration, i.e. transforming the correlated surrogate outputs into uncorrelated linear combinations for the evaluation of the sampling criterion. Further investigation in this direction will be pursued in future work.

Acknowledgements

This work was supported by the Czech Science Foundation under project No. 26-20456S. The development of theoretical background was supported by the Ministry of Education, Youth and Sports of the Czech Republic under project No. LUAUS24260.

References

Blatman, G. and B. Sudret: 2011, 'Adaptive sparse polynomial chaos expansion based on least angle regression'. *Journal of Computational Physics* **230**(6), 2345–2367.

Borri, A. and E. Speranzini: 1997, 'Structural reliability analysis using a standard deterministic finite element code'. *Structural Safety* **19**(4), 361–382.

Choi, S.-K., R. V. Grandhi, R. A. Canfield, and C. L. Pettit: 2004, 'Polynomial Chaos Expansion with Latin Hypercube Sampling for Estimating Response Variability'. *AIAA Journal* **42**(6), 1191–1198.

Echard, B., N. Gayton, and M. Lemaire: 2011, 'AK-MCS: An active learning reliability method combining Kriging and Monte Carlo Simulation'. *Structural Safety* **33**(2), 145–154.

Fajraoui, N., S. Marelli, and B. Sudret: 2017, 'Sequential Design of Experiment for Sparse Polynomial Chaos Expansions'. *SIAM/ASA Journal on Uncertainty Quantification* **5**(1), 1061–1085.

Hadigol, M. and A. Doostan: 2018, 'Least squares polynomial chaos expansion: A review of sampling strategies'. *Computer methods in Applied Mechanics and Engineering* **332**, 382–407.

Jones, D. R., M. Schonlau, and W. J. Welch: 1998, 'Efficient global optimization of expensive black-box functions'. *Journal of Global optimization* **13**(4), 455–492.

Marelli, S. and B. Sudret: 2018, 'An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis'. *Structural Safety* **75**, 67–74.

Novák, L.: 2022, 'On distribution-based global sensitivity analysis by polynomial chaos expansion'. *Computers & Structures* **267**, 106808.

Novák, L., M. Vořechovský, V. Sadílek, and M. D. Shields: 2021, 'Variance-based adaptive sequential sampling for Polynomial Chaos Expansion'. *Computer Methods in Applied Mechanics and Engineering* **386**, 114105.

Papaioannou, I., C. Papadimitriou, and D. Straub: 2016, 'Sequential importance sampling for structural reliability analysis'. *Structural Safety* **62**, 66–75.

Schöbi, R., B. Sudret, and S. Marelli: 2017, 'Rare Event Estimation Using Polynomial-Chaos Kriging'. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* **3**(2).

Schueremans, L. and D. V. Gemert: 2005, 'Benefit of splines and neural networks in simulation based structural reliability analysis'. *Structural Safety* **27**(3), 246–261.

Shields, M. D., K. Teferra, A. Hapij, and R. P. Daddazio: 2015, 'Refined stratified sampling for efficient Monte Carlo based uncertainty quantification'. *Reliability Engineering & System Safety* **142**, 310–325.

Sudret, B.: 2008, 'Global sensitivity analysis using polynomial chaos expansions'. *Reliability Engineering & System Safety* **93**(7), 964–979.

Vořechovský, M.: 2022, 'Reliability analysis of discrete-state performance functions via adaptive sequential sampling with detection of failure surfaces'. *Computer Methods in Applied Mechanics and Engineering* **401**, 115606.

Xiu, D. and G. E. Karniadakis: 2002, 'The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations'. *SIAM Journal on Scientific Computing* **24**(2), 619–644.