A RECOMMENDATION SYSTEM FOR FIRST-ORDER NOAB DESIGNS WITH MULTIPLE PERFORMANCE MEASURES

Zachary C. Little, The Perduco Group, Beavercreek, OH 45431, zach.little@theperducogroup.com
Jeffery D. Weir, Department of Operational Sciences, AFIT, jeffery.weir@afit.edu
Raymond R. Hill, Department of Operational Sciences, AFIT, raymond.hill@afit.edu
Brian B. Stone, Department of Operational Sciences, AFIT, brian.stone.1@us.af.mil
Jason K. Freels, Department of Systems Engineering and Management, AFIT, jason.freels@afit.edu
Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson Air Force Base, Ohio 45433, USA

ABSTRACT

The construction of nearly orthogonal-and-balanced (NOAB) designs is examined for full first-order models in a framework that is informed by the algorithm selection problem for multiple design performance measures and various design size and imbalance settings. Based on a randomly-generated set of large decision spaces, the choice of design size drives the changes in other design performance measures, with decision space features found to impact the measures as well. In this multi-objective setting, prediction of design performance within the framework consistently results in the recommendation of designs that perform well over an entire weight space in addition to designs for specific weights.

Keywords: space-filling design, meta-model, desirability function, synthesized efficiency

INTRODUCTION

Large decision spaces for complex, black-box systems often cannot be exhaustively explored, requiring space-filling experimental designs with possibly mixed factors (i.e., quantitative and qualitative with different numbers of levels). Such designs allow for the construction of meta-models to efficiently represent system responses, and the nearly orthogonal-and-balanced (NOAB) mixed-factor designs are a popular approach for these situations. Orthogonality allows for examination of individual factors separately and can be measured by the maximum absolute pairwise correlation of design matrix columns, denoted by ρ_{map}. An orthogonal design has $\rho_{map} = 0$, while a *nearly orthogonal* design has $\rho_{map} \leq 0.05$. The first-order NOAB designs are created to ensure near orthogonality between first-order model terms (i.e., main effects). A design is considered *nearly balanced* when the maximum imbalance for all factor columns, δ, is close to zero, which ensures that all levels for a factor are represented nearly equally. A construction method is developed for first-order NOAB designs in [1], though beyond a suggested range for the number of design points there exists a need for greater knowledge of design performance for different design sizes and other construction parameter settings. With design matrix columns constructed sequentially by solving various mixed-integer linear programs (MILP), there are many possible parameter settings that could be examined to determine how to create the "best" performing design for a specific study. The framework of an algorithm selection problem can aid in such understanding by examining different parameter settings in the design construction method for a number of different decision space problems. The aim is to accurately predict design performance to allow for efficient design selection and construction. This knowledge will also allow for the development of a recommendation system that accounts for multiple design performance measures of possible interest to an analyst.

Meta-learning was developed to understand learning algorithm performance for classification problems, and developments in meta-learning from many different fields have been generalized and presented in a unified framework in [2] that considers the algorithm selection problem as a learning problem. Rice formalized the algorithm selection problem in [3], where the abstract model (Figure 1) is comprised of a problem space P, feature space F, algorithm space A, and performance space Y, with the algorithm selection problem stated as follows:

> "*For a given problem instance* $x \in P$*, with features* $f(x) \in F$*, find the selection mapping* $S(f(x))$ *into algorithm space* A*, such that the selected algorithm* $\alpha \in A$ *maximizes the performance mapping* $y(\alpha(x)) \in Y$." [2]

The selection of a mapping function S is also an algorithm selection problem. Though the algorithm space A of interest will be a set of parameter settings for design construction, previous work in meta-learning for meta-model selection and other selection problems from [4] [5] [6] [7] can inform a model-based S that accurately predicts design performance measures based on meta-features from the problem space (i.e., set of decision spaces). The process permits the ranking of algorithms (i.e., parameter settings) and can lead to automated learning.

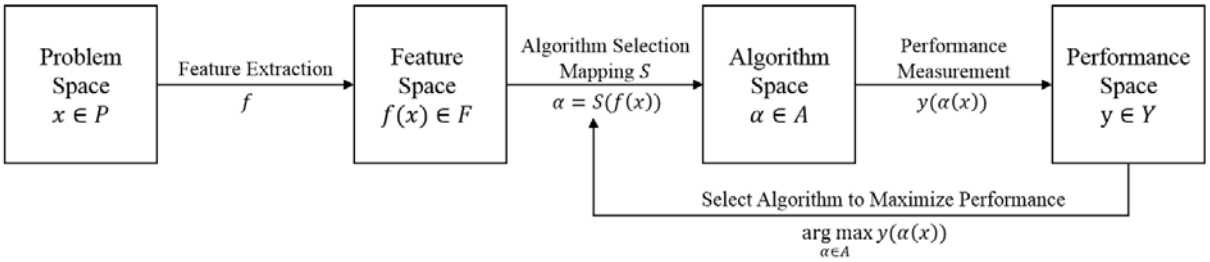

Figure 1. Diagram of Rice's model [2] [3] [4]

Design evaluation and comparison for when multiple performance measures are of interest are discussed, which will lead to how the performance space Y is defined. The algorithm selection problem for first-order NOAB design construction is then presented, with computational results for design performance as well as prediction performance of the resulting recommendation system provided.

METHODOLOGY

Experimental Design Evaluation and Comparison

With respect to design performance measures, focus is placed on low experimental cost (the number of design points, n, for a design matrix X) as well as good model parameter estimation and prediction accuracy. The average and maximum *unscaled prediction variance*, $\text{UPV} = x^{(m)\prime}(X'X)^{-1}x^{(m)}$, over all possible design points m are examined, as in [8]. When it is infeasible to compute the exact average or maximum UPV over a large decision space, an estimate is calculated using a Monte Carlo approach for up to ten million points from the design/decision space. In order to consistently estimate UPV values, all constructed designs for the same decision space problem are compared using the same sampling of points. For good parameter estimation, the *D-criterion*, $|X'X|^{1/p}$, from [9] is used. Due to finding similar overall trends for the average and maximum UPV measures, only maximum UPV is used as a design performance measure in the framework due to the greater variability seen over design choices. In this multi-objective setting, the aim is to minimize n and maximum UPV, while also maximizing the D-criterion.

The measures of various objectives should have the same scale in order to be comparable, so linear, one-sided *desirability functions* [10] are used for each of the criteria, with lower and upper limits set relative to the available designs [11]. A common approach for forming an overall desirability function for m objectives is the multiplicative function $D = \prod_{i=1}^{m} {d_i}^{w_i}$, for individual desirability scores d_i and weights w_i where $\sum_{i=1}^{m} w_i = 1$. The multiplicative function ensures that no individual objective scores too low. *Synthesized efficiency* (SEff), defined as $D(X, w_1, \dots, w_m) / \max_{X^*} D(X^*, w_1, \dots, w_m)$ for design X, is used to examine how X compares to the top performing design for various weightings $(w_1, \dots, w_m)$ of overall desirability [11]. These techniques for design evaluation and comparison are used to obtain aggregate measures for the performance space.

Algorithm Selection Problem

The ***problem space*** consists of 30 randomly-generated decision spaces (Figure 2) with between 8 and 20 factors overall, where categorical factors have between 3 and 7 levels and discrete factors have between 2 and 12 levels. Previous work in decision support efforts for portfolio selection inform the decision spaces having multiple factors of the same type with the same number of levels. Note that continuous factors in NOAB designs are a special case of discrete factors with n levels equally spaced between zero and one.

Problem	Factor (number of levels)																			
21	**6**	**6**	**6**	**6**	**4**	**4**	**4**	**4**	12	12	12	8	8	6	4	4	4	4	4	2
13	**6**	**6**	**6**	**6**	**6**	**3**	**3**	11	10	10	10	10	10	6	6	6	6	6	2	2
12	**7**	**7**	**6**	**3**	**3**	12	12	12	12	11	11	11	11	11	5	5	5	5	2	2
15	**6**	**5**	**5**	**5**	**5**	12	11	11	10	10	10	8	7	6	6	6	5	3	3	3
19	**6**	**4**	**4**	**4**	**4**	12	11	11	10	10	9	9	9	9	7	4	4	4	4	4
20	**7**	**7**	**7**	**7**	**7**	**6**	**6**	11	11	11	11	6	6	6	6	2	2	2	2	
10	**7**	**7**	**7**	**6**	**6**	**6**	**6**	**6**	**3**	12	10	10	9	7	7	7	7	7		
25	**3**	**3**	**3**	11	11	10	10	10	7	7	5	5	5	5	5	3	3	2		
1	**7**	**7**	**7**	**7**	**5**	**5**	11	11	11	11	11	10	7	7	6	6				
29	**7**	**6**	**6**	**6**	**6**	**6**	12	9	9	7	5	5	5	5	2	2				
22	**6**	**6**	9	9	9	9	6	6	6	6	6	5	2	2	2	2				
18	**7**	**7**	9	9	9	9	9	7	7	6	6	6	6	6						
17	**6**	**6**	**6**	**4**	**4**	**4**	**4**	**4**	6	6	6	6	5							
4	**7**	**6**	**6**	**6**	**6**	10	10	10	10	9	5	5	5							
23	**5**	**5**	**5**	**3**	9	9	9	9	5	5	5	5	5							
26	**4**	**4**	**4**	10	10	10	10	10	3	3	2	2	2							
11	**6**	**6**	**6**	**6**	**6**	**4**	**4**	6	6	6	6	6								
8	**6**	**6**	**6**	**3**	**3**	11	11	11	9	4	4	3								
3	**5**	**5**	**5**	8	8	3	3	3	3	3	2	2								
6	**6**	**6**	**6**	10	9	9	7	4	4	4	4	4								
2	**6**	**6**	10	10	10	10	10	5	5	5	5	5								
24	**4**	**4**	8	8	8	7	7	6	6	6	6	6								
30	**4**	**4**	11	11	11	11	6	6	5	5	5	2								
27	**6**	**6**	7	7	7	7	7	6	6	6	2									
7	**3**	9	9	9	9	7	4	4	4	3	3									
28	10	10	9	5	3	3	3	3	3	2	2									
14	**5**	**5**	**5**	**4**	**4**	11	11	11	11	2										
16	**7**	**7**	**3**	**3**	10	10	8	5	5											
9	**4**	**3**	12	8	8	8	8	8	4											
5	**7**	**4**	**4**	8	8	6	6	6												

Categorical

Discrete

Figure 2. Generated problem set of decision spaces

The ***algorithm space*** is comprised of combinations of $m = 2, 3, \dots, 10$ and maximum allowed imbalance $\delta^* = 0.05, 0.10, 0.15, 0.20, 0.25$. The smallest balance-feasible design size $n = m_{bf} \cdot s \geq m \cdot s$ is attempted for each choice of m where s is the number of design matrix columns, so it is possible that multiple (m, δ^*) combinations result in a single combination of (m_{bf}, δ^*). Larger δ^* values allow for

greater imbalance and typically smaller values of balance-feasible n. Each MILP considers the set of design matrix columns for a single factor and is permitted up to two attempts of 30 seconds each to satisfy near orthogonality ($\rho_{map} \leq 0.05$). However, resulting designs with $\rho_{map} > 0.05$ are also recorded for better prediction of design performance. It is possible that some smaller designs may not be able to achieve near orthogonality, yet may have acceptable ρ_{map} depending on the particular study. Larger designs may require more run time in the MILP solver to achieve near orthogonality due to greater computational requirements.

The ***feature space*** includes 24 meta-features with the goal of sufficiently describing each decision space problem: the number of factors for each factor type (discrete and categorical) as well as statistical measures of the number of levels for each factor type, to include minimum, mean, maximum, Q1, median, Q3, sum, standard deviation, skewness, and kurtosis. The product of all numbers of levels (i.e., full factorial design size) and least common multiple of all numbers of levels are also included as meta-features.

The ***performance space*** is multi-objective where the aim is to minimize design size n and estimated maximum UPV, while maximizing the parameter estimation measure D-criterion. As previously discussed, linear desirability functions of the three measures form an overall multiplicative desirability, with weights given to each individual desirability. The entire weight space $\{(w_1, w_2, w_3) | \sum_{i=1}^{3} w_i = 1\}$ is sampled using a 5,000-point space-filling mixture design. While multiplicative desirability for a specific set of weights can be informative, designs that are robust to weightings can also be found by examining average and minimum synthesized efficiencies (SEffs) over the weight space. With respect to overall desirability, average SEff, and minimum SEff over the weight space, the relative performance of the top five predicted designs is compared with that of the actual top performing design and Spearman's rank correlation coefficient is used to compare the actual and predicted rankings.

A model-based approach examines a set of possible mappings S from the parameter settings and meta-features to each of the performance measures, where the meta-model providing the smallest root mean square error (RMSE) for each measure is selected. The meta-models considered include artificial neural networks (ANN) [12] [13] [14], classification and regression trees (CART) [15], multivariate adaptive regression splines (MARS) [16], Gaussian processes (GP) with linear, polynomial, and radial kernels [17] [18] [19], random forests (RF) [20], and support vector machines (SVM) with linear, polynomial and radial kernels [21] [22]. Each meta-model uses the standard parameter grid search settings from the R package *caret*. The training and test instances are important for determining the meta-model S, so all observations for the problem to be predicted are held out from the training data. In order to reduce bias in design performance predictions, 10-fold cross-validation is used where the training data is randomly partitioned so that all designs for the same problem instance will exist in either the training or validation set for each of the folds.

COMPUTATIONAL RESULTS

First-order NOAB Design Performance

Design construction is implemented in MATLAB R2015a using CPLEX V12.6.1 to obtain MILP solutions. Over the 30 decision space problems, there are 1,304 constructed designs in total, resulting from distinct combinations of m_{bf} and δ^* parameters. In Figure 3, there are clear trends in D-criterion as well as average and maximum UPV estimates over the true design size n and relative size m_{bf}. For designs of the same size, those requiring fewer columns tend to be more desirable for each design performance

measure. The relative design size m_{bf} appears to have a strong relationship with average UPV, while designs with fewer columns tend to have higher maximum UPV for designs of the same relative size. It is clear that the choice of relative design size m is important as well as the number of columns s in the design matrix. The number of columns is comprised of defined meta-features, since each discrete factor is represented by a single column and each categorical factor with ℓ levels is represented by $\ell - 1$ columns when using effect coding.

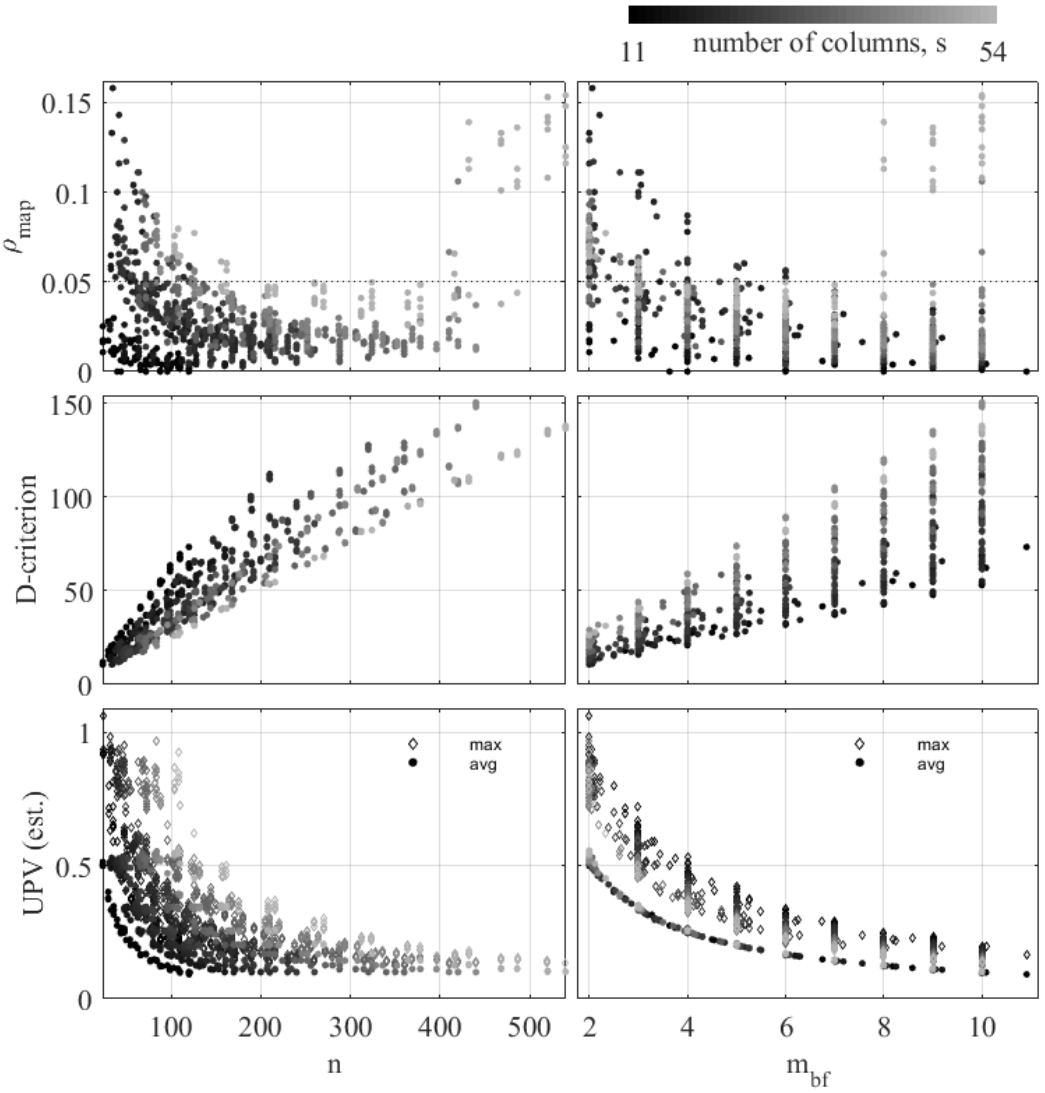

Figure 3. First-order NOAB design performance

Only 181 of 1,304 constructed designs are found to not be nearly orthogonal (ρ_{map} > 0.05), yet 26 larger designs (with $m_{bf} \geq 8$) can be constructed with near orthgonality when the MILP solver is permitted 60 seconds rather than 30 seconds per attempt (not shown in Figure 3). This is consistent with the overall trend for ρ_{map} as well as the idea that larger designs have greater computational requirements. When provided enough time in construction, it appears that larger designs will generally result in sufficient ρ_{map}. The remaining 155 smaller designs that do not satisfy near orthogonality suggest that if small n is of the greatest concern to an analyst, even for problems requiring a small number of design matrix columns, they should examine whether the resulting ρ_{map} is sufficient for their particular problem.

Prediction Performance of Recommendation System

Design size n is predetermined by each choice of m and δ^* (and thus, m_{bf}) using the balance-feasibility test from [1]. For prediction of D-criterion, SVM with a polynomial kernel results in the smallest RMSE over all 30 training sets, with no other meta-model providing similarly small RMSE. For maximum UPV, RF provides the smallest RMSE for all 30 training sets with an average RMSE of 0.0225, while MARS provided the second best average of 0.0260. Figure 4 shows the actual versus predicted values of D-criterion and maximum UPV as well as their respective desirability scores for all 1,304 designs. The desirability scores for D-criterion are scaled relative to the designs found for each problem, which appear to resolve some of the bias that exists for a small number of problems.

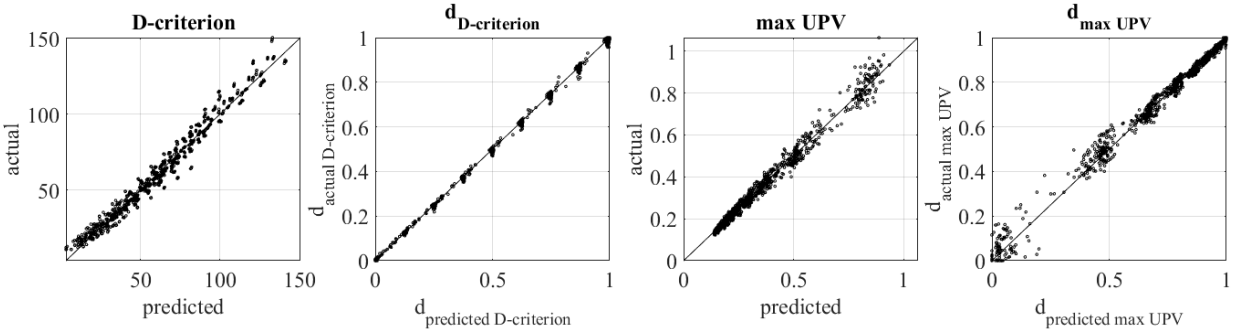

Figure 4. Actual by predicted design performance

Table 1. Top-k relative performance and Spearman's correlation coefficient

			Robust Selections		Multiplicative Desirability Over Weight Space	
Over Problem Space			**avg SEff**	**min SEff**	**min**	**avg**
Top-k Relative Performance	1	min	0.9817	0.9809	0.8143	0.9823
	2		0.9906	0.9868	0.8813	0.9888
	3		0.9913	0.9913	0.9601	0.9920
	4		0.9914	0.9922	0.9601	0.9943
	5		0.9914	1.0000	0.9601	0.9974
	1	avg	0.9959	0.9966	0.9758	0.9965
	2		0.9978	0.9976	0.9916	0.9983
	3		0.9982	0.9989	0.9926	0.9991
	4		0.9985	0.9994	0.9940	0.9995
	5		0.9987	1.0000	0.9964	0.9998
Spearman's correlation coefficient		min	0.9613	0.9469	0.8475	0.9681
		avg	0.9764	0.9732	0.9652	0.9872

The larger residuals for high maximum UPV (low desirability) occur when design size n is small (high desirability), causing a small region of the weight space to have lower top-k relative performance and Spearman's correlation coefficient when examining the multiplicative overall desirability (Table 1). Otherwise, the top-k relative performance and Spearman's correlation coefficient are satisfactory for both robust design recommendations using average SEff and minimum SEff as well as multiplicative desirability for specific weights. For example, if we examine the top-1 relative performance for multiplicative desirability, the worst case (minimum) over both the weight space and problem space gives

0.8143, while the worst-case average over the 30 problems is 0.9753 and the worst-case average over the weight space is 0.9823. Though parameters associated with the most desirable designs will change over the weight space, common selections for m_{bf} across all problems are 6 and 7 for high average SEff (often near 0.89) and 6 for high minimum SEff (often near 0.5). Increasing δ^* generally relaxes balance constraints to achieve smaller n, and thus, m_{bf}.

For a single decision space in this set of problems, the best and worst case for computation time required to construct designs for all (m_{bf}, δ^*) combinations are approximately 2 and 14 hours, respectively. For the recommendation system, building meta-models for D-criterion and maximum UPV on existing design data requires roughly 30 seconds when using the respective mappings of SVM with polynomial kernel and RF. Constructing a single, recommended design within this problem space needs only between 3 and 19 minutes. It is clear that the developed framework and resulting recommendation system allow for efficient selection and construction of first-order NOAB designs.

CONCLUSIONS AND FURTHER RESEARCH

This work shows it is possible to accurately predict first-order NOAB design performance measures for various design sizes and maximum allowed imbalance settings. These predictions permit a recommendation system that can provide both robust selections in the form of designs that have high average and maximum SEff over the weight space as well as designs that perform well for specific weights. For the 30 decision space problems considered, larger designs are generally more desirable with respect to good model parameter estimation as well as low prediction variance. Decision spaces with more design matrix columns tend to need more design points to achieve performance similar to other, smaller decision spaces.

Little et al. have derived extensions to the original first-order construction method to allow for the creation of second-order NOAB designs (i.e., near orthogonality includes two-way interactions and quadratic effects) [23], which may be examined in a similar framework. The second-order extensions also allow for an examination of *NOAB resolution IV* screening designs, in contrast to the first-order NOAB, or *NOAB resolution III* designs, that are the focus of this work. Additionally, future work could examine the computational requirements of these approaches based on the decision space of interest, whether by changing the allowed run time or implementing other stopping criteria for the MILP solver. A comparison with computer-generated optimal designs is also warranted for a large number of decision spaces with multiple performance measures of interest.

ACKNOWLEDGMENTS

This research is in support of the Simulation and Analysis Facility (SIMAF), Air Force Life Cycle Management Center, Simulation and Analysis Division (AFLCMC/XZS). The views expressed in this paper are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

REFERENCES

[1] H. Vieira Jr., S.M. Sanchez, K.H. Kienitz, M.C.N. Belderrain, Efficient, nearly orthogonal-and-balanced, mixed designs: an effective way to conduct trade-off analyses via simulation, J. Simul. 7 (2013) 264–275. doi:10.1057/jos.2013.14.

[2] K.A. Smith-Miles, Cross-disciplinary perspectives on meta-learning for algorithm selection,

ACM Comput. Surv. 41 (2008) 1–25. doi:10.1145/1456650.1456656.
[3] J.R. Rice, The algorithm selection problem, Adv. Comput. 15 (1976) 65–118. doi:10.1016/s0065-2458(08)60520-3.
[4] C. Cui, M. Hu, J.D. Weir, T. Wu, A recommendation system for meta-modeling: A meta-learning based approach, Expert Syst. Appl. 46 (2016) 33–44. doi:10.1016/j.eswa.2015.10.021.
[5] M.A. Muñoz, Y. Sun, M. Kirley, S.K. Halgamuge, Algorithm selection for black-box continuous optimization problems : A survey on methods and challenges, Inf. Sci. (Ny). 317 (2015) 224–245. doi:10.1016/j.ins.2015.05.010.
[6] G. Loterman, C. Mues, Selecting accurate and comprehensible regression algorithms through meta learning, in: Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW 2012, 2012: pp. 953–960.
[7] A.L.D. Rossi, A.C.P. de L.F. de Carvalho, C. Soares, B.F. de Souza, MetaStream: A meta-learning based method for periodic algorithm selection in time-changing data, Neurocomputing. 127 (2014) 52–64.
[8] C.M. Anderson-Cook, C.M. Borror, D.C. Montgomery, Response surface design evaluation and comparison, J. Stat. Plan. Inference. 139 (2009) 629–641. doi:10.1016/j.jspi.2008.04.004.
[9] L. Lu, C.M. Anderson-Cook, T.J. Robinson, Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier, Technometrics. 53 (2011) 353–365. doi:10.1198/tech.2011.10087.
[10] G. Derringer, R. Suich, Simultaneous optimization of several response variables, J. Qual. Technol. 12 (1980) 214–219.
[11] R.H. Myers, D.C. Montgomery, C.M. Anderson-Cook, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, John Wiley & Sons, Hoboken, NJ, 2016.
[12] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408.
[13] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133.
[14] D.O. Hebb, The Organization of Behavior: A Neuropsychological Theory, John Wiley & Sons, New York, 1949.
[15] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA, 1984.
[16] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1991) 1–141.
[17] G. Matheron, Principles of geostatistics, Econ. Geol. 58 (1963) 1246–1266.
[18] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer simulation experiments, Stat. Sci. 4 (1989) 409–423. doi:10.1214/ss/1177012413.
[19] T.J. Santner, B.J. Williams, W.I. Notz, The Design and Analysis of Computer Experiments, Springer, New York, 2003.
[20] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.
[21] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Syst. 9 (1997) 155–161.
[22] S.M. Clarke, J.H. Griebsch, T.W. Simpson, Analysis of support vector regression for approximation of complex engineering analyses, J. Mech. Des. 127 (2005) 1077–1087.
[23] Z.C. Little, J.D. Weir, R.R. Hill, B.B. Stone, J.K. Freels, Second-order Extensions to Nearly Orthogonal-and-balanced (NOAB) Mixed-factor Experimental Designs, Work. Pap. (2017).
[24] Z.C. Little, J.D. Weir, R.R. Hill, B.B. Stone, J.K. Freels, A Recommendation System for First-order Nearly Orthogonal-and-balanced (NOAB) Designs, PhD Colloquium Extended Abstract, Proceedings of the 2017 Winter Simulation Conference, W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds., (2017).