Analyzing Capabilities of Latin Hypercube Designs Compared to Classical Experimental Design Methods

Andreas Fischbach, Jörg Stork, Martin Zaefferer, Sebastian Krey, Thomas Bartz-Beielstein

SPOTSeven Lab, Dept. of Comp. Sci. and Eng. Sci.
TH Köln
E-Mail: firstname.lastname@th-koeln.de
www.spotseven.de

Introduction

Design of experiments (DOE) has proven to be a useful tool to optimize process or production output [1]. A design specifies which values for input parameters of an experiment are to be chosen to reach a desired output or gather a maximum amount of information.

Many authors recommend space-filling and non-collapsing designs for deterministic computer experiments [2, 3]. *Latin hypercube designs* (LHDs) are non-collapsing by default due to their creation rules. The space-filling property can be fulfilled by maximizing the minimum Euclidean distance of points. Those designs are especially useful for fitting Kriging models [4]. But it has not been proven, if they are the best choice [2]:

> *In sum, it has not been demonstrated that LHDs are superior to any designs other than simple random sampling (and they are only superior to simple random sampling in some cases).*

Based on the experimental analysis of Santner [2], this work analyses the behavior of different LHD types (space-filling and non-space-filling) used to fit Kriging metamodels and compare their properties and performance to classical design of experiment methods.

Designs can be evaluated with regards to optimality. For a given model, information based statistical criteria like D-optimality (maximize the determinant of the information matrix), A-optimality (minimize the trace of the

inverse of the information matrix) and E-optimality (maximize minimum eigenvalue of the information matrix) can be calculated.

Searching for an optimal design even with the search space limited to a subclass of all designs is still difficult [5]. Often, a search for an optimal design is stopped before convergence, to avoid spending too much time on design generation. So, in this work the approach is to randomly draw designs instead of creating optimized designs with much effort. Afterwards the drawn designs are evaluated and compared with optimized designs to determine, which properties and criteria a design must fulfill to be adequate for fitting Kriging metamodels.

This work compares model dependent properties like the mentioned optimality criteria with spatial criteria like minimum and average Euclidean interpoint distance. The goal of this paper is to find answers to the following research questions:

- Which criteria must a design fulfill to enable fitting high quality Kriging metamodels?
- Does the design type have a significant impact on the model quality?
- How many points should a good design consist of (regarding the tradeoff between cost intensive number of experiments and the model quality)?

In this work, the designs' properties and their effect on building Kriging models will be observed using a simple test function. This enables a very transperent and easy to reproduce comparison. Also, sufficiently accurate models may be build with a very limited number of design points.

Methods

Optimality Criteria

The optimality criteria observed in this work can be divided in two groups. The first contains model dependent criteria and the second spatial criteria with no reference to an underlying (surrogate) model. The model-dependent

criteria are computed based on the information matrix of each design corresponding to a simple first order linear regression model [6].

Model-dependent Criteria

Consider the linear model in matrix form, $\mathbf{Y} = \beta\mathbf{X} + \epsilon$, where the vector ϵ contains the n random errors (n denotes the sample size). The elements of the error vector, ϵ_i, are assumed to be independent and identically normally distributed with mean zero and *error variance* σ_ϵ^2. The vector of unknown coefficients, β, can be estimated via least squares methods as $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}$. Its covariance matrix is $\text{var}(\hat{\beta}) = \sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$. The square roots of the diagonal entries of this matrix are the standard errors of the β_i's. The diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$ are called the relative variances of the β_i's, $v_i = \text{var}(\hat{\beta})/\sigma_\epsilon^2$. The v_i's indicate how large the variances of the estimated model parameters compared to the error variance σ_ϵ are. The inverse of the covariance matrix $\text{var}(\hat{\beta})$ is called the *information matrix* for the model parameter vector $\hat{\beta}$, i.e., $\mathbf{M} = 1/\sigma_\epsilon^2\mathbf{X}^T\mathbf{X}$. The information matrix is a measure of the information about the factor effects that is contained in the design. The information matrix measures the information on factor effects that is contained in the design.

An experimental design that maximizes the determinant of the information matrix, $|\mathbf{M}|$, is called *D-optimal*. The minimization of $|(\mathbf{X}^T\mathbf{X})^{-1}|$ is equivalent to the maximization of $|(\mathbf{X}^T\mathbf{X})|$.

A-optimality minimizes the trace of the inverse of the information matrix and is often used as a criterion for designs with equal D-optimality.

E-optimality maximizes the smallest eigenvalue of the information matrix. The aim of E-optimality is to minimize the maximum variance of all possible normalized linear combinations of parameter estimates. E-optimal designs minimize the maximum axis of the confidence ellipsoid of estimators, namely, E-optimal designs minimize the maximum eigenvalue of the covariance matrix of estimators.

An overview of these criteria is given in [7]. The relationship between these three optimality criteria is illustrated in [8].

Spatial Criteria

The spatial criteria analyzed in this work are the Cartesian pairwise point distances of a design. Let m denote the problem dimension and $d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^{m} |X_{ik} - X_{jk}|^p\right)^{1/p}$, with $p \in \{1, 2\}$, be the *distance* between two *design points* X_i and X_j. A *maximin* distance design maximizes the minimum pairwise point distance, i.e., $d_{\min} = \min_{1 \leq i,j \leq n, i \neq j} d(X_i, X_j)$. Additional optimization criteria for LHDs are discussed in [5]. *Minimax* distance designs are defined in a similar manner [9].

Quality Criteria

In the following we will analyze the model quality for each design fitting a Kriging model and measure the impact of the mentioned design properties, namely

- minimum distance: $d_{\min}$,
- average distance: d_{avg},
- optimality: A-, D-, and E.

Note, not every criterion from this list is used as an optimality criterion during the design generation. For example, the average distance will be used only to visualize and analyze results. Our analysis is based on results obtained by Santner [2].

Design types

To answer the previously stated questions several experimental designs will be generated. The below described designs are divided in different types, namely *Maximin LHD*, *Degenerated LHD*, *Optimized LHD* and *Uniform*. All types have in common that their designs are created based on pseudo random number sampling. Therefore experiments using these designs are repeated multiple times and statistically analyzed to lead to robust results.

The *Maximin LHD* uses pairwise Cartesian distances of the design points to maximize the minimum distance and thus lead to a space filling design.

Out of 1,000 randomly created LHDs the one with the maximum minimum distance will be taken.

Degenerated LHD tries to compute a non space-filling Latin hypercube design to be able to analyze the importance and necessity of the space-filling property. For this type the design out of 1,000 randomly generated LHDs with the minimum sum of the pairwise Euclidean distances will be drawn. An *Optimized LHD* is computed using the simulated annealing algorithm to optimize an LHD via the Φp criterion, which is linked to the minimum distance criteria [10]. Regarding computation time, especially for large designs, this design type is computationally very cost intensive.

The various LHDs are compared to a simple random sampling approach. Here, design points are sampled from a uniform distribution, hence this approach is denoted *Uniform*. The comparison of *LHD* and *Uniform* approaches provides a simple benchmark. Furthermore, it allows to determine the necessity of the space-filling and non-collapsing property of the LHDs

To get an idea if and how classical designs are also applicable for computer experiments and if they are able to outperform LHDs when fitting Kriging-models, *Full Factorial Designs* (FFD) with appropriate number of design points and additional center points are also included in the experiments.

Metamodel

Kriging (or Gaussian Process Regression) is a frequently used surrogate-model. It is an excellent predictor of smooth, continuous problem landscapes. Moreover, it provides an uncertainty estimate of its own prediction, which can be used to calculate the *Expected Improvement* (EI) of a candidate solution. EI is one great benefit of Kriging models and is for example used in the Efficient Global Optimization algorithm introduced by Jones et al. [11] to balance exploitation and exploration in a Kriging-based optimization process.

The Kriging implementation used in the experiments is taken from the SPOT R-Package[1]. This implementation is based on earlier code by

[1] See `https://cran.r-project.org/package=SPOT`

Forrester et al. [12], who also provide a very comprehensive description of Kriging and Kriging-based optimization.

A Kriging model with default parameters is used. The quality of a design is computed by evaluating the *Root Mean Squared Error* (RMSE) of a derived Kriging model. For that purpose, the residuals of the model on an equidistant grid are computed.

Experiments

The described experiments are based on work done by Santner in 2003 [2]. Santner compared the behavior of a space-filling and evenly spread LHD with a degenerated LHD with design points at the diagonal axis of a unit square.

Objective Function

The experiments were conducted using the objective function defined in Santner [2]:

$$y(x_1, x_2) = \frac{x_1}{1 + x_2} \qquad (x_1, x_2) \in [0, 1] \times [0, 1]. \tag{1}$$

This objective function is relatively easy to model (see Figure 1). It allows to focus on the behavior of the design according to its properties instead of the modeling method itself. The surface of the function is almost plain showing a linear trend and there is an interaction between x_1 and x_2.

Setup

30 random instances of the previously described design types are drawn for design sizes from $n = 5$ to $n = 12$ design points. For $n = 5$ and $n = 9$ a FFD design is added resulting in a total number of 962 simulation experiments (30 $\times$ number of design types $\times$ number of different design sizes + 2 FFD experiments). In each experiment the properties of the used design are computed (pairwise Euclidean distances and A-, D- and E-optimality).

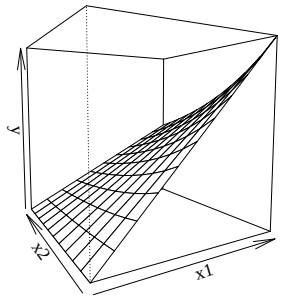

Figure 1: Objective Function used in all Experiments.

For each design, a Kriging model of the objective function is fitted using the function values, y_i, at the n design points X_i. Afterwards the model fit is used to predict values at a pre-defined grid over the region of interest from 0.05 to 0.95 in each dimension with a distance of 0.1. These values are referred to as $\hat{Y}_i$. This results in $l = 100$ prediction values that are used to compute the difference to the objective function values at the grid points. With these values the RMSE as defined in Equation 2 is computed.

$$\text{RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i - \hat{Y}_i)^2}. \tag{2}$$

To answer the question if there is a correlation between the model based optimality criteria of the designs and the quality of the fitted Kriging models, the optimality criteria are computed using the design information matrix $\mathbf{M}$ with regard to a linear regression model. The model contains terms for each main effect x_1 and x_2 and the interaction term $x_1 * x_2$.

Results

An overview of the results from the experiments is given in Figure 2. The boxplot shows the median value and the upper and lower quartiles of the

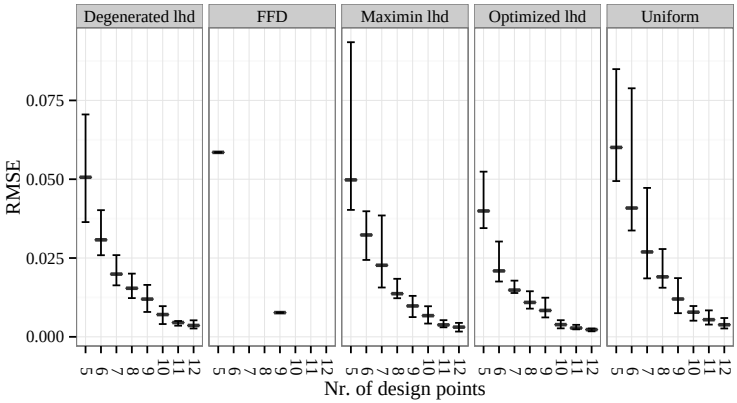

Figure 2: Development of the model error per design type by increasing number of design points.

RMSE of a set of experiments. The values are grouped by the design type and the number of design points. The column FFD shows only two results due to the fact that the Full Factorial Designs with center points are deterministic. It can be seen, that model error and variance of error decrease with increasing number of design points. Although there are further improvements to be expect when design points are added, a convergence for each type is tendentially recognizable. Further increasing of the design size would not lead to a significantly improved model performance.

In the FFD column the experiment results of the two single experiments for five and nine points are shown. The five point FFD performance is similar to the median of the other five point designs. On the other hand, the nine point FFD's performance seems to be better than the median performance of the other types.

The performance differences of the design types can best be seen with a minimum number of design points, say five to seven points. The median RMSE of the Uniform designs are clearly worse and the variance is largest. It is somehow surprising that the *Degenerated LHD* seems to perform at the same level compared to the others, except of *Uniform*. In sum it can be stated, that it is not so easy to consistently create bad performing

non-space-filling LHDs.

Table 1: Analysis of Variance of experiment results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
type	3	0.0232	0.0077	19.1479	4.63e-12	***
nrDesignPoints	1	0.3035	0.3035	751.1904	< 2e-16	***
minDist	1	0.0004	0.0004	0.8743	0.350014	
avgDist	1	0.0056	0.0056	13.8912	0.000205	***
dOptimality	1	0.0115	0.0115	28.5166	1.16e-07	***
aOptimality	1	0.0001	0.0001	0.3022	0.582629	
eOptimality	1	0.0087	0.0087	21.6129	3.81e-06	***
Residuals	950	0.3838	0.0004			

To conclude if the design type among some or all of the designs properties have a significant impact on the model error an ANOVA is performed. The two experimental results of the *Full Factorial Designs* for five and nine design points are removed in advance.

Table 1 shows the results of the ANOVA. The analysis shows that the type has a significant effect on the model error. Besides the type it can be seen, that the number of design points (as expected), the average Euclidean distance, D-optimality and E-optimality have a significant impact at a 99% significance level. Regarding the mean of squares the number of design points has obviously the largest impact on the model error. The other significant parameters seem to influence the model error at almost the same level.

To further analyze and compare the different used design types and to answer the question which design type works best a Tukey's honest significance test is performed to present a pairwise comparison of the different design types. The results are shown in Figure 3. The *Degenerated LHD* and *Maximin LHD* perform on an equal level. Also the *Optimized LHD* and the *Degenerated LHD* perform not significantly different, the same holds for the *Optimized LHD* and *Maximin LHD*. All LHD types are superior to *Uniform* in performance at a 95% confidence level.

One stated question is which properties of the described experimental designs have a significant impact on the model quality. The ANOVA results

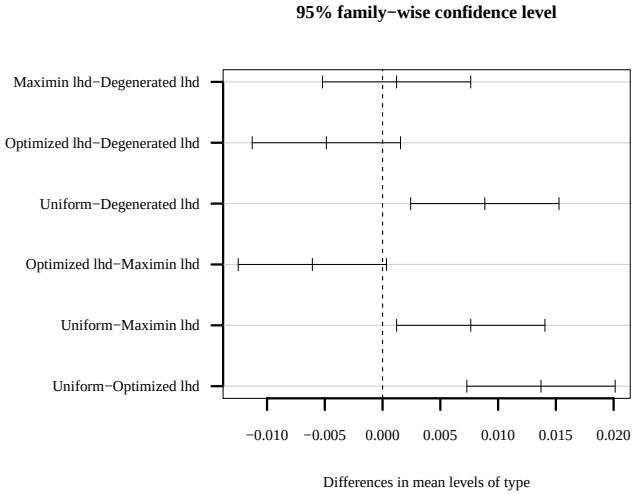

Figure 3: Pairwise differences of model error for LHD types at 95% confidence level. This figure illustrates,e.g., that Uniform Designs (random sampling) is worse compared to Optimized LHD, because the corresponding confidence interval does not contain zero. It is positive, so the difference in the RMSE values is positive, too.

(see Table 1) shows the significance of the design properties. Figures 4 to 6 show the development of the resulting RMSE for increasing values of the minimum distance, D-optimality and A-optimality grouped by the number of design points. The solid black line shows a trend of the RMSE for the parameter, computed via local smoothing of the collected data. In these results, the two experiments with the *FFD* are included.

Regarding the ANOVA results, a look at the trend of the *minimum distance* parameter (see Figure 4) also shows a decreasing RMSE by increasing the minimum distance, but not consistently for all number of design points and additionally the effect is decreasing with increasing design size.

D-optimality is often analyzed together with *A-optimality*. If two or more designs with an equal value for D-optimality are found, the experimenter often chooses the design with the best value for *A-optimality*. Figure 5

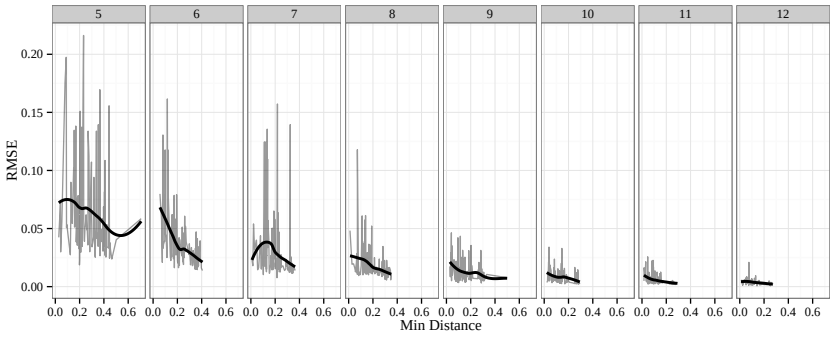

Figure 4: Effect of minimum distance property $d_{\min}$ on model error grouped by number of design points.

clearly shows the positive effect of *D-optimality* on the model RMSE. In contrast to the general trend, rather high RMSE values can be observed for the rarely occuring large D-optimality values.

It also can be seen that with larger designs the value range for the determinant (printed at a logarithmic scale) is decreasing and lower values occur less frequent. Figure 6 shows the development of the *A-optimality*, which is desired to be maximized. A positive impact of A-optimality on the RMSE can be recognized, especially for seven and nine design points. This is observed despite of the fact that the earlier described ANOVA evaluated the impact of A-optimality to be not significant.

Regarding the researched question how many points a design should consist of at least, tt can be seen that for five design points the variance of the results is quite large. Slightly increasing the design size up to seven points leads to better models and reduced variance in the RMSE (see Figure 2). The *Optimized LHD* seems to be the most robust design type.

The best five designs out of 120 for a design size of $n = 5$ and $n = 9$ points with their ranked properties are given in Table 2 and 3. For a design size of five it can be seen that the Optimized LHD leads to good ranked values (third and sixth best values) for minimum distance but not automatically to the best RMSE. The best design, a *Degenerated LHD* has no property resulting in a top ten rank, but good values for D-optimality

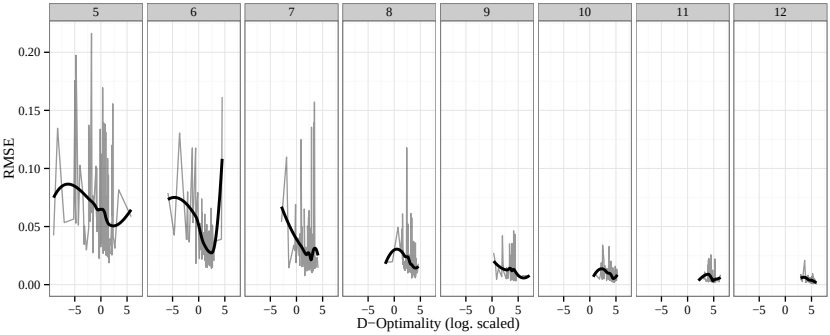

Figure 5: Effect of D-optimality, $|\mathbf{M}|$, on model error grouped by number of design points.

(17th), A-optimality (26th) and E-optimality (35th). For nine design points at least the best design is an *Optimized LHD*, but the RMSE values of the following designs differ not much. The best designs properties yield good ranks (7th and 8th) for D-, A- and E-Optimality. The third best design is to be mentioned because the computed properties are all ranked around the 90th place out of 120. So it seems worth to not only focus on one single

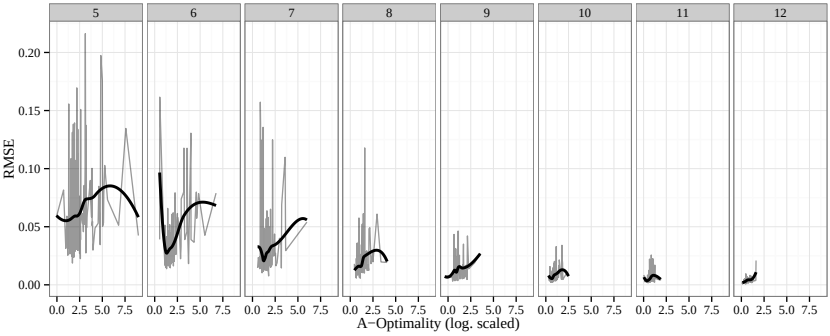

Figure 6: Effect of A-optimality, trace($\mathbf{M}$), on model error grouped by number of design points.

Table 2: Best performing designs and ranked properties for n=5

type	RMSE	minDist	avgDist	D-opt	A-opt	E-opt
Degen. lhd	0.0187	87	41	17	26	35
Maximin lhd	0.0225	86	104	84	78	78
Optim. lhd	0.0237	3	56	24	38	44
Optim. lhd	0.0244	6	30	27	20	21
Degen. lhd	0.0253	46	82	31	29	31

Table 3: Best performing designs and ranked properties for n=9

type	RMSE	minDist	avgDist	D-opt	A-opt	E-opt
Optim. lhd	0.0031	16	38	7	7	8
Maximin lhd	0.0035	94	15	35	36	38
Maximin lhd	0.0036	98	97	92	87	84
Degen. lhd	0.0038	62	8	14	8	7
Degen. lhd	0.0041	55	30	43	59	61

criterion to optimize when looking for suitable designs for fitting Kriging metamodels.

Conclusion

This work analyzes the behavior of different design of experiment methods and the design properties including optimality criteria. The focus is on determining whether these properties and criteria have an influence on the quality of a Kriging model. The research covers the questions of the impact of the design type, the significance of model dependent optimality criteria and spatial criteria and the performance development of the designs by increasing design size.

The experiment results show that LHDs are superior to designs created by simple random sampling, especially at a small number of design points. *Optimized LHDs* are more robust and perform equal to *Maximin LHDs* and *Degenerated LHDs*. Hence the usage of them can be advantageous if

the situation allows to afford the computation time. If computation time is an important aspect *Maximin LHDs* should be sufficient.

It seems that the average distance of the design points can be more important than the minimum distance. Further studies are required to verify that thesis. A positive effect of model dependent optimality criteria (for first order linear regression models) can be recognized, although the effect is not significant for all computed criteria. Regarding the results of the FFD experiments, it seems interesting to combine the abilities of standard designs with LHDs.

- **Which criteria must a design fulfill to enable fitting high quality Kriging metamodels?**
  A general answer to this question was not found, but maximizing the minimum interpoint distance does not automatically lead to the best designs. Similar to E-optimality, the smallest eigenvalue of a correlation matrix of a Kriging model may be maximized. This may provide an interesting approach towards generating optimal designs for Kriging models.

- **Does the design type have a significant impact on the model quality?**
  Simple random sampling lead to significant worse designs. But it is also not easy to generate consistently bad performing LHDs, especially for larger designs. Surprisingly the different examined LHD types perform all on an equal level. This might be due to the simplistic experiment setup. A more sophisticated experimental analyses might yield deeper insights to effects of different design types.

- **How many points should a good design consist of (regarding the tradeoff between cost intensive number of experiments and the model quality)?**
  The answer to this question depends on the structure and complexity of the objective function and the expected costs of an experiment. In this case designs consisting of at least seven points lead to a significant decrease in the variance and the median values of the model error. For design sizes larger than seven points the improvement of the model quality becomes less significant.

Interesting further studies include the reproduction of the results on different more complex test functions created by a randomized function generator. The work should be extended to find statistical robust criteria for optimal designs for Kriging models.

Acknowledgements

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Energie unter den Förderkennzeichen KF3145101WM3 und KF3145103WM4 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

References

[1] W. Kleppmann, *Versuchsplanung.* Hanser, 2013.

[2] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments.* Berlin, Heidelberg, New York: Springer, 2003.

[3] B. Husslage, G. Rennen, E. van Dam, and D. den Hertog, "Space-filling latin hypercube designs for computer experiments," *Optimization and Engineering*, vol. 12, no. 4, pp. 611–630, 2011.

[4] J. P. C. Kleijnen, *Design and analysis of simulation experiments.* New York NY: Springer, 2008.

[5] R. Jin, X. Du, and W. Chen, "The use of metamodeling techniques for optimization under uncertainty." Journal of Structural & Multidisciplinary Optimization (in press), 2005.

[6] D. C. Montgomery, *Statistical Quality Control.* Wiley, 2008.

[7] H. Bandemer and A. Bellmann, *Statistische Versuchsplanung.* Teubner, 1994.

[8] T. Takeuchi and H. Sekido, "An Approximate Approach to E-optimal Designs for Weighted Polynomial Regression by Using Tchebycheff Systems and Orthogonal Polynomials," *ArXiv e-prints*, Mar. 2013.

[9] S. Crary, "Design of computer experiments for metamodel generation," *Analog Integrated Circuits and Signal Processing*, vol. 32, no. 1, pp. 7–16, 2002.

[10] M. D. Morris and T. J. Mitchell, "Exploratory designs for computational experiments," *Journal of Statistical Planning and Inference*, vol. 43, pp. 381–402, Feb. 1995.

[11] D. Jones, M. Schonlau, and W. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, pp. 455–492, 1998.

[12] A. Forrester, A. Sobester, and A. Keane, *Engineering Design via Surrogate Modelling*. Wiley, 2008.