High-Dimensional Model-Based Optimization Based on Noisy Evaluations of Computer Games

Mike Preuss[1], Tobias Wagner[2], and David Ginsbourger[3]

[1] Technische Universität Dortmund, Chair of Algorithm Engineering (LS 11)
Otto-Hahn-Str. 14, Dortmund, D-44227, Germany
mike.preuss@tu-dortmund.de

[2] Technische Universität Dortmund, Institute of Machining Technology (ISF)
Baroper Str. 301, Dortmund, D-44227, Germany
wagner@isf.de

[3] University of Bern, Institute of Mathematical Statistics and Actuarial Science
Sidlerstr. 5, Bern, CH-3012, Switzerland
david.ginsbourger@stat.unibe.ch

Abstract. Most publications on surrogate models have focused either on the prediction quality or on the optimization performance. It is still unclear whether the prediction quality is indeed related to the suitability for optimization. Moreover, most of these studies only employ low-dimensional test cases. There are no results for popular surrogate models, such as kriging, for high-dimensional ($n > 10$) noisy problems. In this paper, we analyze both aspects by comparing different surrogate models on the noisy 22-dimensional car setup optimization problem, based on both, prediction quality and optimization performance. In order not to favor specific properties of the model, we run two conceptually different modern optimization methods on the surrogate models, CMA-ES and BOBYQA. It appears that kriging and random forests are very good modeling techniques with respect to both, prediction quality and suitability for optimization algorithms.

Keywords: Computer Games, Design and Analysis of Computer Experiments, Kriging, Model-Based Optimization, Sequential Parameter Optimization, The Open Racing Car Simulator.

1 Introduction

Over the last 15 years, the use of surrogate-model-assisted optimization approaches has obtained a high popularity in almost all application areas [12,13,17,23]. Within this period, the research on model-based optimization has mainly focused on low-dimensional problems and noise-free evaluations. In particular, kriging has been shown to be well-suited for modeling deterministic data of computer experiments (design and analysis of computer experiments, DACE [21]) with low or moderate input dimension $n \in [1, 10]$. In the modeling and optimization of practical problems, however, e. g., in the computer games community,

Y. Hamadi and M. Schoenauer (Eds.): LION 6, LNCS 7219, pp. 145–159, 2012.

high-dimensional parameter spaces and noisy responses have to be considered. Consequently, the modern kriging models of DACE have been enhanced to cope with noisy data in recent years [5, 6, 11]. For high-dimensional data, however, almost no results of kriging-based modeling approaches have been reported.

In this paper we thus investigate how these kriging variants and other popular surrogate modeling techniques can assist in optimizing a 22-dimensional problem from the domain of computer games - the car setup optimization problem based on the open racing car simulator (TORCS). The response to be modeled is the distance obtained by a racing car with a specific car setup encoded by the input parameters. Based on a short evaluation time on TORCS with a quasi-random starting point on the track, this response is very noisy. The analysis and evaluation of the surrogate models is two-fold. First, their global prediction qualities based on the initial design are evaluated. Then, the capability of the models to guide and tune the optimization [19] is assessed by performing a global optimization on the model and compare the predicted optimum with the quality of the evaluation on TORCS (one-step approach). Almost all previous studies using a one-step approach have focused only on one of these aspects – the prediction quality or the results of a model-based optimization. Based on the combined analysis, some important questions can be addressed:

1. Is the prediction quality a good indicator for the optimization capability of a surrogate model?
2. Are certain surrogate models particularly well suited for high-dimensional noisy problems?
3. Can the successful results of kriging-based optimization approaches be transferred to higher dimensions and noisy data?

In the following section, the basic principles of the considered surrogate models are described. The car setup optimization problem and TORCS are briefly summarized in section 3. The two main sections of the papers address the prediction quality and the optimization results obtained by the different surrogate models. In the final section 6, the results are summarized, conclusions are drawn, and an outlook on future research topics is given.

2 Surrogate Models

For almost all real-world applications, the evaluation of parameter vectors is time-consuming and/or expensive, e. g., because a finite-element analysis, a computational fluid dynamics calculation or a real-world experiment have to be performed. In these cases, a model-based approach is often used. Here, we focus on one-step approaches. Based on an initial design of the problem parameters, a model is fitted which is then used as a surrogate for the actual experiment, e. g., the parameter vector resulting in the optimal model prediction is directly used as a solution or the model is used as a surrogate for tuning optimization algorithms in order to use the tuned variant on the actual problem [19]. For both kinds of applications, the surrogate model should

1. be as close to the true response as possible (prediction quality), and
2. reflect the characteristics of the optima of the true response surface (model optimization).

In the following subsections, some popular surrogate models are described and discussed. Due to the extremely high number of approaches, we restrict our description to the models considered in our experiments. More methods and additional details to the presented approaches can be found in Hastie et al. [9].

2.1 First Order Response Surface

In the first order (linear) response surface model (LM), the relationship between the control variables $\mathbf{x}_i$ and the corresponding observations y_i is described by

$$y_i = \mathbf{x}_i \beta + \varepsilon_i. \tag{1}$$

Equation 1 is set up for each pair of parameter vector $\mathbf{x}_i$ and observation y_i ($i = 1, \ldots, N$) in the initial design. The least-squares estimate $\hat{\beta}$ of the coefficients β is then calculated as the solution to the corresponding system of linear equations [10, p. 11]. With this $\hat{\beta}$, equation 1 can be used for prediction of unknown parameter vectors $\mathbf{x}$.

2.2 Generalized Additive Model

The Generalized Additive Model (GAM) [8] replaces the linear form of equation 1 by a sum of smoothing functions for single parameters $\beta + \sum s_j(x_j)$ ($j = 1, \ldots, k$), where an iterative algorithm is employed to decide about the important variables x_j and the corresponding smooth functions s_j. Contrary to the first order response surface (LM), the GAM also allows nonlinear smoothing functions to be specified. The employed R package `GAM`[1] supports local polynomial regression and smoothing splines.

2.3 Random Forest

Random forests [2] consist of huge ensembles (typically 500 or more) of decision trees, whereby each of them is trained on a randomly chosen subset of the available observations. The prediction of the random forest is then computed as the average of the predictions of the individuals trees. Random forests are usually used for classification, but also regression can be realized by implementing regressing decision trees, as done in the R package `randomForest`[2].

[1] `http://cran.r-project.org/web/packages/gam/index.html`

[2] `http://cran.r-project.org/web/packages/randomForest/index.html`

2.4 Kriging

Kriging is a surrogate model originated from geosciences [4] which has become popular in the DACE [21] and machine learning [20] communities. In ordinary kriging, the response of interest can be considered as one realization of a random variable $Y(\mathbf{x}) = \mu + Z(\mathbf{x})$, where $\mu \in \mathbb{R}$ is an intercept used for centering the stationary zero-mean Gaussian process (GP) Z. Z depends on a covariance kernel of the form $(\mathbf{x}, \mathbf{x}') \to D^2 : k(\mathbf{x}, \mathbf{x}') \mapsto \sigma^2 r(\mathbf{x} - \mathbf{x}'; \psi)$ for a correlation function r with parameters ψ.

The predictions of the kriging model can be obtained by taking the conditional expectation $m(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})|Y(\mathbf{x}_i) = y_i]$ of Y based on the N current pairs of parameter vectors $\mathbf{x}_i$ and observations y_i. Consequently, $m(\mathbf{x})$ is also denoted as the kriging mean. It provides a prediction for each observation $\mathbf{x}$ by enhancing the constant trend using the correlation to the existing observations. It thus explicitly uses the information of each observation. For an efficient evaluation, the kriging mean can be computed in closed form

$$m(\mathbf{x}) = \widehat{\mu} + \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1}(\mathbf{y} - \widehat{\mu}\mathbf{1}), \tag{2}$$

using the observations $\mathbf{y} = (y_1, \ldots, y_n)^T$, the covariance matrix of the experiments $\mathbf{K} = (k(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}))$, the covariance vector $\mathbf{k}_n(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^T$ of $\mathbf{x}$ and the existing design points, and the maximum likelihood estimation of the trend

$$\widehat{\mu} = \frac{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{K}^{-1} \mathbf{1}}.$$

For noisy evaluations $\widetilde{Y}_i := Y(\mathbf{x}_i) + \varepsilon_i$, the GP is conditioned based on a sum of random variables – one following a GP and one for the noise. Assuming independence between the random variables as well as between different realizations of the noise, the kriging mean can still be computed using equation 2, only the intercovariance matrix $\mathbf{K}$ is replaced by $\bar{\mathbf{K}} = \mathbf{K} + \tau^2 \mathbf{I}$ at every occurrence. The additional term τ^2 denotes the noise variance which is only added for identical observations. In the case of heterogeneous noise variances, i.e., $var(\varepsilon_1) = \tau_1^2 \neq \ldots \neq var(\varepsilon_N) = \tau_N^2$, $\mathbf{K}$ is replaced by $\bar{\mathbf{K}} = \mathbf{K} + diag([\tau_1^2 \ldots \tau_n^2])$. Contrarily to the noiseless case, these models do not interpolate the noisy observations.

The choice of the covariance kernel and its parameters determines the shape (smoothness, modality) and the flexibility of the response surfaces predicted by the kriging model. In this paper, two popular kernels implemented in the R package `DiceKriging`[3] are considered:

1. the Gaussian kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left[-\sum_{j=1}^{n} \left(\frac{x_j - x_j'}{\theta_j}\right)^2\right] \tag{3}$$

[3] `cran.r-project.org/web/packages/DiceKriging/index.html`

2. the Matérn kernel with $\nu = 5/2$:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \prod_{j=1}^{n} \left[1 + \sqrt{5} D_j + \frac{5}{3} D_j^2\right] \exp\left[-\sqrt{5} D_j\right],\ D_j = \frac{|x_j - x'_j|}{\theta_j} \quad (4)$$

Both kernels depend on a set of parameters, σ^2 and $\{\theta_1, \ldots, \theta_d\}$, which are often referred to respectively as *process variance* and *ranges*. They have to be fitted based on the available evaluations, for which we use maximum-likelihood estimation in the experiments.

3 Car Setup Optimization Problem

The car setup optimization problem originates from a competition held at the EvoStar 2010 conference[4]. It is based on the open source car racing simulator (TORCS)[5] which is used as simulation engine for the evaluations. The task in this competition is to find a near optimal setting for the 22 car parameters listed in Table 1. Performance is measured by the track distance covered within this time frame. In order to avoid handling different parameter ranges within the optimization, all parameters are scaled to the interval [0, 1] by the interface.

Table 1. The 22 car setup optimization parameters of the EvoStar 2010 competition and their original ranges, taken from [3]

parameter	section	name	unit	min	max
1	gearbox/gears/2	ratio	SI	0	5
2	gearbox/gears/3	ratio	SI	0	5
3	gearbox/gears/4	ratio	SI	0	5
4	gearbox/gears/5	ratio	SI	0	5
5	gearbox/gears/6	ratio	SI	0	5
6	rear wing	angle	deg	0	18
7	front wing	angle	deg	0	12
8	brake system	front-rear brake repartition	SI	0.3	0.7
9	brake system	max pressure	kPa	100	150000
10	front anti-roll bar	spring	lbs/in	0	5000
11	rear anti-roll bar	spring	lbs/in	0	5000
12	front left-right wheel	ride height	mm	100	300
13	front left-right wheel	toe	deg	-5	5
14	front left-right wheel	camber	deg	-5	-3
15	rear left-right wheel	ride height	mm	100	300
16	rear left-right wheel	camber	deg	-5	-2
17	front left-right suspension	spring	lbs/in	0	10000
18	front left-right suspension	suspension course	m	0	0.2
19	rear left-right suspension	spring	lbs/in	0	10000
20	rear left-right suspension	suspension course	m 0	0.2	
21	front left-right brake	disk diameter	mm	100	380
22	rear left-right brake	disk diameter	mm	100	380

In the competition, a time frame of one million *tics* of 20 *ms* each was allowed as budget for the optimization algorithm. The algorithm can distribute the available time arbitrarily between different settings, i. e., each evaluation can take as

[4] `http://cig.ws.dei.polimi.it/?page_id=103`

[5] `http://torcs.sourceforge.net/`

long as desired. For comparable results, however, the evaluation time should be fixed. The evaluations are made in a row while the game is running, where short breaks are required in order to brake down the car to a standstill. Therefore, different parts of the track are used for measuring the performance, which is a major source of the noise in the evaluations. Recent parameter studies [16] have shown that below a certain limit of around 250 tics (5 s), measured values become so noisy that they are unsuitable for optimization. Longer evaluations spent the budget more quickly so that Kemmerling recommends evaluations of 2000 tics (40 s), resulting in only 500 evaluations of the simulator. In this time, around one third of the Suzuka F1 track (wheel-2 in TORCS) can be covered. This track is shown in Fig. 1. It combines many challenges, such as high speed parts and different curve types, and is, thus, used for evaluations in this paper.

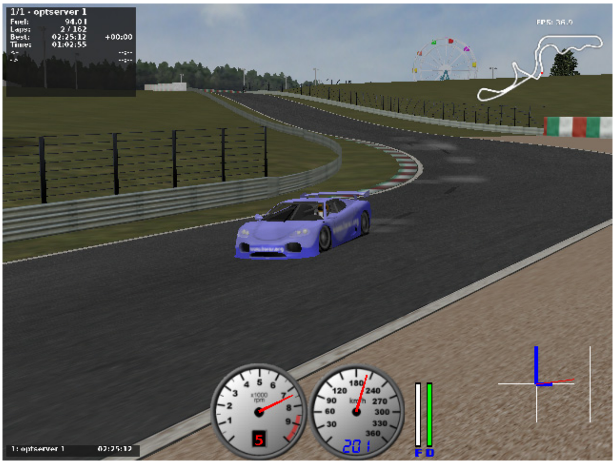

Fig. 1. Screenshot of TORCS, in which the solution obtained from the model-based optimization on the (Matérn covariance) kriging model is driving the reference track (Suzuka F1). A minimap of the track is shown in the top right corner.

Summarizing, the car setup optimization problem can be regarded as a high-dimensional noisy practical problem with a very limited budget of evaluations. Consequently, it is hard to solve[6]. In the computer games context, such problems appear whenever an implemented, parameterizable component (as a car driving bot) must be adapted to other components, be they provided by the user or procedurally generated [16]. As known from the formula 1 races, the optimal setup will change whenever one of the components (car, driver, track) is modified. Thus, the results of the EvoStar2010 competition where other tracks then the Suzuka F1 were considered cannot be directly compared to our setup.

[6] See also the results of the EvoStar 2010 competition at `http://www.slideshare.net/dloiacono/car-setup-oprimization-competition-evostar-2010`

4 Prediction Quality of the Models

In the first part of the experimental analysis, the prediction quality of the surrogate models on the car setup problem is evaluated. The results build the basis for the combined analysis in the next section.

Setup. In order to evaluate the prediction quality of the models, sets for the training and the validation of the surrogate model had to be prepared. The budget for the training set was chosen according to the specification of the car setup competition, where 20000 s of running the simulator had been allowed. We chose an evaluation time of 40 s which resulted in a size of 500 design points in the training set. As mentioned in the previous section, the evaluations of a car setup were noisy. In order to allow the effect of the accuracy of an evaluation to be considered, two different training sets were used, one having 125 mutual design points with 4 replications $\mathbf{X}_{125,4}$ and one having 500 mutual design points without replications $\mathbf{X}_{500,1}$. For validation, a larger and more accurate set of $N = 440$ randomly distributed design points with 20 replications was employed. The root mean squared error $RMSE = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2}$ was used as performance measure, where $\hat{y}_i$ denotes the prediction of the model for the i-th design point and $\bar{y}_i$ is the mean of the 20 observations in the validation set. For some parameter vectors, the damage of the car exceeds a specified threshold. In these cases, the output of the simulator is not the distance reached by the car, but a high penalty encoding the damage. In order to not deteriorate the quality of the models by integrating discontinuities and different scales, these values were removed from the training and validation sets. If a subset of the repeats of a parameter vector is penalized, only the remaining results were used for the computation of the mean performance.

For kriging, we analyzed the effect of the covariance kernel and the use of the estimated variances of set $\mathbf{X}_{125,4}$ in a heterogeneous kriging model. More specifically, we consider the Gaussian and the Matérn kernel with $\nu = 5/2$ and the homogeneous (single τ^2) and heterogeneous (vector of τ_i^2) formulation of kriging for noisy observations, as presented in section 2.4. The other surrogate models have been run with their standard parameters. For LM and GAM, the set $\mathbf{X}_{125,4}$ was tested with and without using variance-depending weights $w_i = 1/var(y_i)$ of the observations, where the latter approach results in a weighted least squares fit.

Pre-experimental Planning. The 22 design variables of the car setup problem directly result in 23 model parameters (θ_j and σ) in the likelihood optimization. In order to avoid a deterioration of the kriging results based on a bad fit of the internal parameters, the optimization of the kriging parameters was analyzed before the experiments. Accounting for the multimodality of the log likelihood, the global optimization strategy of the DICEKriging package based on the Genetic Optimizer Using Derivatives (RGenOUD)[7] was considered. It was observed that

[7] `http://cran.r-project.org/web/packages/rgenoud/index.html`

the standard parameters, population size $P = 20$, wait generations $W = 2$, maximum generation limit $L = 5$, and the generation in which the gradient-based refinement is performed for the first time (BFGS burnin) $B = 0$ did not robustly find the maximum of the log likelihood for the large data set $\mathbf{X}_{500,1}$. Thus, a small experiment was conducted in preparation for the actual study. Based on a 43 point Latin hypercube design (LHD) [14] of the four RGenOUD parameters, the performance of log likelihood optimization was analyzed. The results of the set $\mathbf{X}_{500,1}$ were defined as the test case. The experiment was conducted for the Gaussian and the Matérn kernel with $\nu = 5/2$. Based on the results and allowing a slightly larger computational budget for ensuring robust results, we recommend to use $(P, W, L, B) = (10, 1, 120, 40)$. The computation time for $\mathbf{X}_{500,1}$ is still below 4 minutes on a 3 GHz PC.

The focus of the experiments reported in the paper is clearly on surrogate models based on or related to kriging. This originates from our experience in using these models. Other popular surrogate models, such as support vector regression and neural networks, were also used in the beginning of the study, but since the results of standard R implementations (e1071 package[8] based on the libsvm[9] and the nnet package[10]) were much worse and we did not manage to appropriately adjust these models, we excluded them from the paper. Nevertheless, our results are based on open R packages and the test sets can be downloaded online[11], which allows a comparison of experts in these areas with our results to be realized.

Task. According to the scope of the experiment, four hypotheses were established:

1. More inaccurate points are better than a few with higher accuracy.
2. The use of the estimated noise variances can improve the results on training set $\mathbf{X}_{125,4}$.
3. The Matérn kernel with $\nu = 5/2$ is superior to the standard Gaussian kernel.
4. Kriging is the superior model with respect to prediction quality.

The first hypothesis was based on former results of kriging on noisy data sets [1] and is generally related to the bias-variance tradeoff in machine learning [9]. The second one was straightforward, but surprises with respect to a bad estimation of the variances or an increase of model complexity might occur. The third one was based on the weaker assumptions of the Matérn kernel compared to the Gaussian with respect to the differentiability of the response surface – twice compared to infinitely often. The fourth one was driven by the hope to transfer the results of kriging in lower dimensions to higher ones. However, this was questionable due to former results in the literature [22].

Results/Visualization. The results of the experiments with regard to the prediction quality are summarized in Table 2.

[8] http://cran.r-project.org/web/packages/e1071/index.html

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[10] http://cran.r-project.org/web/packages/nnet/index.html

[11] http://ls11-www.cs.uni-dortmund.de/rudolph/kriging/applications?&#video_game_data

Table 2. Summary of the results with respect to the prediction quality of the models

method	data set	var. used?	repeats	mean RMSE	std RMSE
LM	$\mathbf{X}_{125,4}$	no	1	0.1835478	-
	$\mathbf{X}_{125,4}$	yes	1	0.2428571	-
	$\mathbf{X}_{500,1}$	no var.	1	0.1699225	-
GAM	$\mathbf{X}_{125,4}$	no	1	0.1346388	-
	$\mathbf{X}_{125,4}$	yes	1	0.1803923	-
	$\mathbf{X}_{500,1}$	no var.	1	0.0976035	-
Random Forest	$\mathbf{X}_{125,4}$	no	10	0.1259052	0.0010350
	$\mathbf{X}_{500,1}$	no	10	0.1066116	0.0007794
Kriging, Gauss	$\mathbf{X}_{125,4}$	no	10	0.1447966	0.0329848
	$\mathbf{X}_{125,4}$	yes	10	0.1283082	0.0000040
	$\mathbf{X}_{500,1}$	no var.	10	0.0935839	0.0000007
Kriging, Matérn	$\mathbf{X}_{125,4}$	no	10	0.1202643	0.0000065
	$\mathbf{X}_{125,4}$	yes	10	0.1283162	0.0273566
	$\mathbf{X}_{500,1}$	no var.	10	0.0937369	0.0000006

Observations. Based on the results of Table 2, the hypotheses can be tested[12]:

1. More inaccurate points are indeed better than a few with higher accuracy. The results of the training set $\mathbf{X}_{500,1}$ are significantly improving the prediction quality for all considered surrogate models.
2. The results concerning this hypothesis show no clear trend. For the LM and the GAM, the results of the weighted least squares fit are worse compared to the standard one. For the Gaussian kernel, the mean prediction quality is improved by using the estimated variances while still being robust with respect to the model fitting. For the Matérn kernel, the mean prediction quality and the robustness of the model fitting decrease with estimated variances. This result, however, is based on the fact that the best model (RMSE ≈ 0.115) is only found in 8 of the 10 repeats. In the two remaining cases, bad models (RMSE > 0.178) are returned which result in the observed decrease in mean performance.
3. The only situation in which the Matérn kernel with $\nu = 5/2$ is indeed superior to the usually applied Gaussian kernel is the one for the set $\mathbf{X}_{125,4}$ with unknown variances. In all other scenarios, no significant results can be obtained with respect to the covariance kernel which is of course also based on the high standard deviation of the Matérn kernel on set $\mathbf{X}_{125,4}$ with estimated variances.
4. Kriging is indeed the superior model with respect to prediction quality. Although the superiority is significant, the improvement over Random Forests (set $\mathbf{X}_{125,4}$) and the GAM (set $\mathbf{X}_{500,1}$) is only small.

[12] Due to the partly deterministic results and very low variances of the stochastic approaches, no additional statistical tests have been performed.

Discussion. The most important effect with respect to the prediction quality is the one of the training set. A diverse set of many inaccurate solutions results in a higher prediction quality for all considered models. This effect may be caused by the diminishing gain of information obtained by replications [11]. In addition, four repeats are not enough for a sensible approximation of the variance corresponding to an observation. This conjecture is also based on the bad results of the weighted least squares approaches.

5 Model-Based Optimization

Related to the questions formulated in the introduction, we now analyze whether the established models are useful for optimization, and if there is a relation to the prediction quality we can exploit to predict this suitability.

Pre-experimental Planning. Based on the results of the previous section, we only focused on models based on the set $\mathbf{X}_{500,1}$. We used a stationary approach in which the surrogate model is not refined. Our focus is on the capability of the model to reflect the characteristics of the true response surface based on the large initial design, e. g., for a one-step optimization approach or for tuning optimization algorithms [19]. In order not to bias the results by the choice of the optimizer, two different optimization strategies were used. The *covariance-matrix-adaptation evolution strategy* (CMA-ES)[13] [7] is a powerful evolutionary algorithm for global optimization, whereas *boundary optimization by quadratic approximation* (BOBYQA)[14] [18] is a modern gradient-free local search strategy for box-constrained optimization.

Task. We want to decide if model quality may be employed as a guideline for choosing a model for optimization, namely by judging these two hypotheses:

1. The prediction quality can be used as an indicator for the suitability for a one-step optimization.
2. The Kriging models offering a high prediction quality are particularly suited for reflecting the characteristics of nonlinear problems and finding its optima.

The first hypothesis expresses the common belief of just considering one of these indicators for assessing surrogate models. The last one was based on the huge number of publications on kriging metamodeling [23].

Setup. In order to evaluate the suitability for optimization, one model of each type was chosen as representative. This choice was made based on the internal quality criterion of the model – log likelihood for kriging, out-of-bag error for random forest. For the other approaches, no variation in the model fitting exists. On each representative, 20 runs of BOBYQA and the CMA-ES were conducted. The obtained local optima were then evaluated on TORCS for 10 times.

[13] `http://cran.r-project.org/web/packages/cmaes/`

[14] `http://cran.r-project.org/web/packages/minqa/`

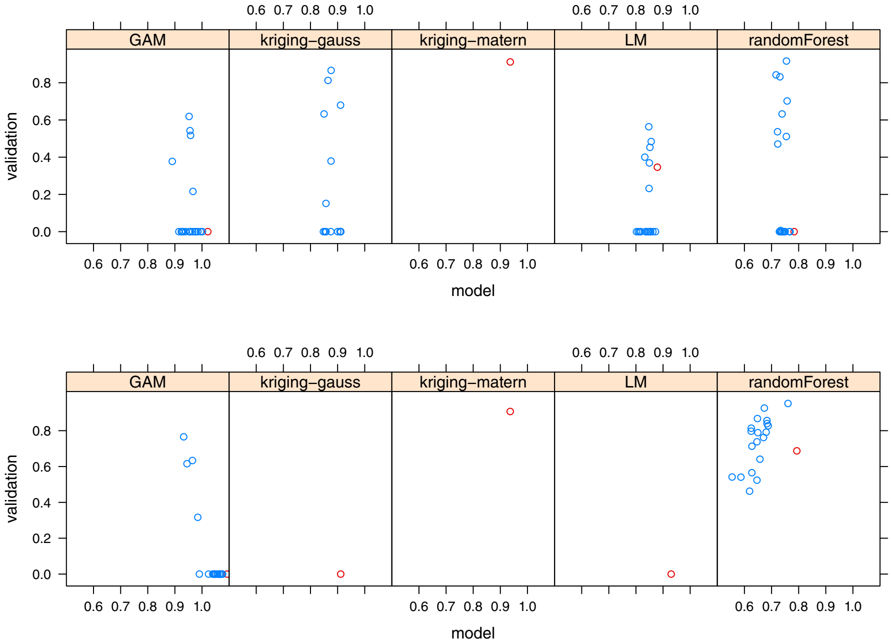

Fig. 2. Comparison of the predicted ('model', x-axis) and actual performance ('validation', y-axis) of the local optima (top: CMA-ES, bottom: BOBYQA.). The potentially global optimum on the model is colored in red.

The resulting values provided the basis for the analysis. If the predicted values of the potential optima also result in locally or globally optimal values on the simulator, the characteristics of the model are well reflected. The assessment of the correlation between the prediction quality and the optimization performance is performed by comparing the RMSE of the representative model and the actual quality of the approximated optima.

Results/Visualization. The results of the optimization on the representative models are shown in Fig. 2. The correlation between the prediction quality and the optimization performance can be assessed based on Fig. 3. It looks conspicuous that a lot of the potential optima on the models result in a validated value of zero. These values are caused by parameter vectors that could not be evaluated properly on TORCS because the damage threshold of the car is exceeded (cf. section 3), making this parameter vector infeasible.

Observations. The first hypothesis can clearly be rejected. Based on Fig. 3, the random forest provides a better optimization performance than Gaussian kriging and GAM, whereas its RMSE is worse. The RMSE can only successfully distinguish between the worst (LM) and the best (Matérn kriging) approach with respect to both indicators.

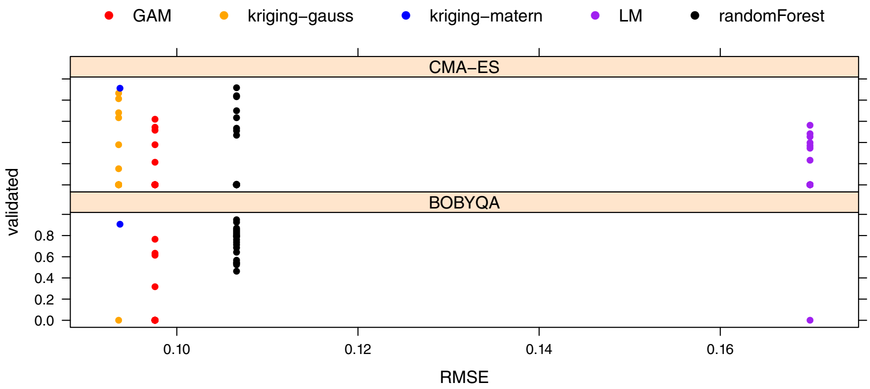

Fig. 3. Correlation between the RMSE of the representative model and the actual averaged distance achieved in TORCS by the approximated optimum parameter setting

With respect to the second hypothesis, no objective results can be seen from Fig. 2. Whereas the CMA-ES on the Gaussian kriging model returned many local optima with an almost equal performance on the model, but high variation on the actual problem, the optimization on the model using the Matérn kernel always resulted in the same optimum which is indeed a good parameter setting. All other nonlinear surrogate models also resulted in a multimodal response surface with different local optima, where the performance variation on the actual problem is huge.

Discussion. The high number of infeasible solutions proposed by LM, GAM, and Gaussian kriging is based on the extrapolation properties of these approaches. Since the parameter vectors of the infeasible solutions are not used for fitting the model, no points for interpolation are provided in these parameter regions. Nevertheless, the assumption of an underlying model (LM and GAM) or the strong differentiability assumed by Gaussian kriging result in local optima within these areas. This is highly undesired for the focused application setting, where only one iteration of model-based optimization is performed, because it may result in an infeasible solution after optimization. In contrast, the Matérn kernel seems to result in a more data-dependent prediction without extrapolation effects.

The difference between the RMSE and the optimization performance may consequently be caused by the different extrapolation properties of the approaches. Because the infeasible values are also removed from the validation set, no predictions in these areas are considered. In addition, the predictions of the random forest seem to be more conservative. Whereas all other modeling approaches predict their local optimal values between 0.8 and 1, the optimal values of the random forest are between 0.6 and 0.8. It seems like the characteristics of the problem are well covered, but the variation of the response values is decreased which results in an increased RMSE.

In order to judge whether the response surfaces reflect the characteristics of the true problem, the true modality of the problem is of interest. In figure 4, we

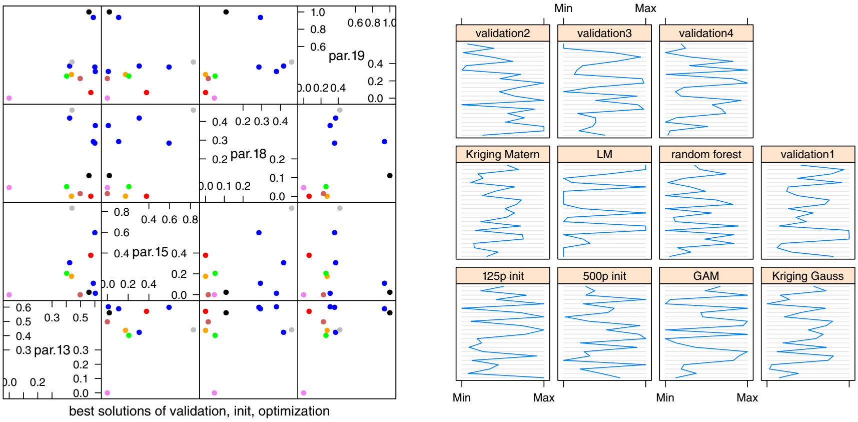

Fig. 4. Visualizations of the 4 best of the 440 solutions of our validation set (blue) together with the best initial solutions from the 125 point design (gray) and the 500 point design (black) and the best solutions obtained from the CMA-ES runs, GAM (red), Kriging Gauss (orange), Kriging Matern (green), LM (violet), random forest (indianred). Left: scatter plot of the 4 most important variables as indicated in [15]. Note the different range for parameter 13 due to infeasible solutions outside this range. Right: parallel coordinate plot over all 22 parameters.

depict the best 4 of the 440 solutions in the validation set (together with the best initials solutions and the best from the optimization runs). The plots clearly show that there are several distinct local optima – even if only the four most important variables according to importance estimations of the GAM model and further studies [15] are considered. There is seemingly not one dominant basin of attraction. Based on this fact, employing the Matérn kriging model that always leads the optimization algorithm to one search space region may become risky. Obtaining different local optima – if they indeed exist – is surely preferable in order to have alternatives for the optimal solution on the model, in case it is infeasible on the actual problem. In particular since the optimization on the surrogate model is very fast compared to an evaluation on the original problem. However, the validated result of the Matérn kriging is extremely good (0.912). The best of the initial points only obtains a validated score of 0.900. The model thus manages to find better solutions. Only four of the space-filling 440 points of the validation set (0.938, 0.928, 0.918, 0.915) are better. For the best validated solution of the random forest (0.917), the situation is similar. However, this solution is not identified as the global optimum of the model (cf. Figure 2).

Summarizing, the second hypothesis cannot be completely accepted. The kriging models either simplify the characteristics of the true problem (Matérn) or result in undesired solutions (Gauss), whereby the Matérn model is surely the better choice. It may just have the empirical information to detect only one of the basins. The multimodality predicted by the Gaussian kernel is also based on undesired extrapolation effects.

6 Conclusion and Outlook

With respect to three questions posed in the introduction, the first question has to be negated. The prediction quality can only roughly distinguish the suitability of a model for optimization. As an answer to the second question, the Matérn kriging model obtained the best results for both, prediction quality and optimization performance. The performance of the approximated optimum is competitive with the best results of the large validation set which required around 20 times more resources to compute. This achievement, however, comes with a simplification of the model characteristics, maybe caused by a smoothing of true response surface. As a consequence, the successful results of kriging-based optimizers could be transferred to higher dimensions and noisy data, although there is a strong dependence on the applied covariance kernel. For a more robust optimization performance, sequential approaches, which refine the model based on the additional evaluations, can be considered [5, 11, 13, 14].

The results obtained in this study are a first step towards a combined analysis of prediction quality and optimization capability on complex practical problems. Their generality is of course questionable due to the restriction to a single problem instance. In the future, automatically generated instances of the car optimization problem with different tracks, cars, and bots and also instances from other related problems can be used to perform a much broader simulation study in order to improve on this drawback. Results of other surrogate modeling approaches on the open test cases defined in this paper can assist in providing guidelines for applications. In any case we would like to emphasize that modeling difficult noisy high-dimensional problems is obviously useful and possible.

Acknowledgments. This paper is based on investigations of project D5 *Synthesis and multi-objective model-based optimisation of process chains for manufacturing parts with functionally graded properties* as part of the collaborative research center SFB/TR TRR 30, which is kindly supported by the Deutsche Forschungsgemeinschaft (DFG). We would like to thank Markus Kemmerling for providing the evaluator code for the car setup optimization competition and Olaf Mersmann for consulting with respect to the `R` implementation of the CMA-ES.

References

1. Biermann, D., Weinert, K., Wagner, T.: Model-based optimization revisited: Towards real-world processes. In: Michalewicz, Z., Reynolds, R.G. (eds.) Proc. 2008 IEEE Congress on Evolutionary Computation (CEC 2008), pp. 2980–2987. IEEE Press, Piscataway (2008)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Cardamone, L., Loiacono, D., Lanzi, P.L.: Car Setup Optimization. Competition Software Manual. Tech. Rep. 2010.1, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy (2010)
4. Cressie, N.: The origins of kriging. Mathematical Geology 22(3), 239–252 (1990)

5. Forrester, A.I.J., Keane, A.J., Bressloff, N.W.: Design and analysis of 'noisy' computer experiments. AIAA Journal 44(10), 2331–2339 (2006)
6. Ginsbourger, D., Picheny, V., Roustant, O., Richet, Y.: Kriging with heterogeneous nugget effect for the approximation of noisy simulators with tunable fidelity. In: Proc. Joint Meeting of the Statistical Society of Canada and the Société Francaise de Statistique (2008)
7. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation 9(2), 159–195 (2001)
8. Hastie, T., Tibshirani, R.: Generalized additive models. Statistical Science 1(3), 297–318 (1986)
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)
10. Horton, R.L.: The General Linear Model. McGraw-Hill, New York (1978)
11. Huang, D., Allen, T.T., Notz, W.I., Zheng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal on Global Optimization 34(4), 441–466 (2006)
12. Jin, Y.: A comprehensive survey of fitness approximation in evolutionary computation. Soft Computing 9(1), 3–12 (2005)
13. Jones, D.R.: A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization 21(4), 345–383 (2001)
14. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13(4), 455–492 (1998)
15. Kemmerling, M.: Optimierung der Fahrzeugabstimmung auf Basis verrauschter Daten einer Autorennsimulation. Diplomarbeit, TU Dortmund (2010) (in German)
16. Kemmerling, M., Preuss, M.: Automatic adaptation to generated content via car setup optimization in torcs. In: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, CIG 2010, pp. 131–138. IEEE (2010)
17. Kleijnen, J.P.C.: Kriging metamodeling in simulation: A review. European Journal of Operational Research 192(3), 707–716 (2009)
18. Powell, M.J.D.: The bobyqa algorithm for bound constrained optimization without derivatives. Tech. Rep. DAMTP 2009/NA06, Centre for Mathematical Sciences, University of Cambridge, UK (2009)
19. Preuss, M., Rudolph, G., Wessing, S.: Tuning optimization algorithms for real-world problems by means of surrogate modeling. In: Proc. 12th Annual Conference on Genetic and Evolutionary Computation (GECCO 2010), pp. 401–408. ACM, New York (2010)
20. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. Springer (2006)
21. Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Statistical Science 4(4), 409–423 (1989)
22. Tenne, Y., Izui, K., Nishiwaki, S.: Dimensionality-reduction frameworks for computationally expensive problems. In: Fogel, G., Ishibuchi, H. (eds.) Proc. 2010 IEEE Congress on Evolutionary Computation (CEC 2010), pp. 1–8. IEEE Press, Piscataway (2010)
23. Wang, G.G., Shan, S.: Review of metamodeling techniques in support of engineering design optimization. Journal of Mechanical Design 129(4), 370–380 (2007)