S. Krajči (ed.): ITAT 2018 Proceedings, pp. 64–71
CEUR Workshop Proceedings Vol. 2203, ISSN 1613-0073, © 2018 Jakub Repický, Martin Holeňa, and Zbyněk Pitra

Automated Selection of Covariance Function for Gaussian Process Surrogate Models

Jakub Repický[1,2] and Zbyněk Pitra[2,3] and Martin Holeňa[2]

[1] Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
[2] Institute of Computer Science, Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
{repicky,martin}@cs.cas.cz
[3] Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Břehová 7, 115 19 Prague 1, Czech Republic

Abstract: Gaussian processes have a long tradition in model-based algorithms for black-box optimization, where a limited number of objective function evaluations are available. A principal choice in specifying a Gaussian process model is the choice of the covariance function, which largely embodies the prior assumptions about the modeled function. Several methods for learning the form of covariance function have been proposed. We report a work in progress in which the covariance function is selected from a fixed set. The goal of covariance function selection is to capture non-local properties of the objective function and derive a more accurate surrogate model. The model-selection algorithm is evaluated in connection with Doubly Trained Surrogate Covariance Matrix Adaptation Evolution Strategy on the Comparing Continuous Optimizers framework. Several estimates of predictive performance, including cross-validation and information criteria, are discussed. Focus is placed on information criteria suitable for nonparametric methods, and two of them are compared experimentally.

1 Introduction

The principle of continuous black-box optimization is finding extrema of real-parameter objective function analytical definition of which is not known. Such functions, often arising, e. g., in engineering design optimization or material science, can only be evaluated empirically or through simulations. Moreover, obtaining function values may be expensive and affected by noise. The goal of finding a global optimum is usually relaxed in favor of finding a good enough solution within as few objective function evaluations as possible.

Evolution strategies, stochastic population-based algorithms inspired by the process of natural evolution, present a popular approach to continuous black-box optimization. The Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [10, 13] is based on adaptation of the key component of the mutation operator (the covariance matrix) according to the historical search steps. The CMA-ES is considered a state-of-the-art continuous black-box optimizer. Nevertheless, considerable improvements in terms of the number of fitness evaluations can be achieved by use of surrogate models, i. e., statistical or machine learning models of the fitness trained on data gathered during the optimization.

A variety of models for the CMA-ES has been investigated, including but not limited to quadratic approximations [14], ranking support vector machines [16], random forests [4] and Gaussian processes (GPs) [4, 19, 25].[1]

Gaussian process (GP) regression is a nonparametric method, meaning the data are assumed to be generated from an infinite-dimensional distribution, i. e., a distribution of functions. In black-box optimization, the distribution of function values conditioned on observed data can be used to derive a criterion for selecting most promising points for evaluation with the (expensive) fitness. As far as we know, the first optimization method utilizing uncertainty modeled by GPs is Bayesian optimization [17]. In this paper, we are going to build upon the more recent Adaptive Doubly Trained Surrogate CMA-ES (aDTS-CMA-ES), which uses a Gaussian process surrogate models for the CMA-ES, although our approach is directly applicable to Bayesian optimization as well.

A Gaussian process is fully specified by a mean function and a covariance function parametrized a by a small number of parameters. In order to distinguish parameters of the mean and covariance functions from the infinite-dimensional parameter vector – the vector of function values – they are referred to as hyperparameters. In statistical works, the mean and covariance functions are chosen by the statistician in a cycle of model building and model checking.

The goal of this work is to lay out a suitable method for learning the form of covariance function for Gaussian processes in black-box optimization with focus on criteria for evaluating candidate covariance functions. The main hypothesis behind this paper is that a GP with a composite form of its covariance function may result in a more accurate approximation of the objective function and, consequently, better performance of the model-assisted optimization algorithm.

[1] An experimental comparison of selected surrogate-assisted variants of the CMA-ES can be found in [3, 20].

Related Work Learning a composite expression of kernel functions for support vector machines by genetic programming was explored in [7].

Hierarchical kernel learning [2] and Additive Gaussian processes [6] are algorithms for determining kernels composed of lower-dimensional kernels.

The goal of Automatic Statistician project [15] is automatic statistical analysis of given data with output in natural language. The algorithm of structure discovery in GP models [5] is a greedy search in the space of composite covariance functions generated by operators of addition and multiplication recursively applied to basis covariance functions.

Up to our knowledge, structure discovery in GP surrogate models for continuous black-box optimization has not yet been investigated. As a first step towards this goal, we perform selection of the best GP model from a model population that we tried to design large enough to capture structure of typical continuous black-box function but still small enough for model selection to be computationaly feasible.

The paper is organized as follows. Section 2 presents ideas behind surrogate models in evolutionary optimization and aDTS-CMA-ES algorithm. Section 3 describes inference and learning in Gaussian process regression models. Section 4 presents the algorithm for selecting the best GP surrogate model. First results from an early stage of experimental evaluation are presented in Section 5. Section 6 concludes the paper.

2 Surrogate-Assisted Evolutionary Optimization

Evolutionary strategies are stochastic search algorithms based on maintaining a population of candidate solutions, usually encoded as real vectors. In each iteration (generation), a population of λ offsprings is generated from a population of μ parents by operators of recombination and mutation. The new population of parents is selected either from the union of offsprings and parents (*plus* selection), or, provided that $\mu \leq \lambda$, from the offsprings exclusively (*comma* selection).

2.1 CMA-ES

Mutation in evolutionary strategies is usually implemented by sampling from a Gaussian distribution, parameters of which play a crucial role in algorithms' convergence. The main idea behind the CMA-ES is self-adaptation of mutation parameters, especially of the covariance matrix. The CMA-ES repeatedly samples from $\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{C})$ and updates parameters σ^2 (overall step-size), $\mathbf{m}$ (the mean) and $\mathbf{C}$ (the covariance matrix) so that likelihood of successful mutation steps increases under new parametrization.

Algorithm 1 aDTS-CMA-ES

Input: λ (population-size), y_{target} (target value), f (original fitness function), α (ratio of original-evaluated points), $\mathscr{C}$ (uncertainty criterion)

1: $\sigma, \mathbf{m}, \mathbf{C} \leftarrow$ CMA-ES initialize
2: $\mathscr{A} \leftarrow \emptyset$ {*archive initialization*}
3: **while** stopping conditions not met **do**
4: $\{\mathbf{x}_k\}_{k=1}^{\lambda} \sim \mathcal{N}\left(\mathbf{m}, \sigma^2\mathbf{C}\right)$ {*CMA-ES sampling*}
5: $f_{\mathscr{M}1} \leftarrow \text{trainModel}(\mathscr{A}, \sigma, \mathbf{m}, \mathbf{C})$ {*model training*}
6: $(\hat{\mathbf{y}}, \mathbf{s}^2) \leftarrow f_{\mathscr{M}1}([\mathbf{x}_1, \ldots, \mathbf{x}_\lambda])$ {*model evaluation*}
7: $\mathbf{X}_{\text{orig}} \leftarrow$select $\lceil \alpha\lambda \rceil$ best points accord. to $\mathscr{C}(\hat{\mathbf{y}}, \mathbf{s}^2)$
8: $\mathbf{y}_{\text{orig}} \leftarrow f(\mathbf{X}_{\text{orig}})$ {*original fitness evaluation*}
9: $\mathscr{A} = \mathscr{A} \cup \{(\mathbf{X}_{\text{orig}}, \mathbf{y}_{\text{orig}})\}$ {*archive update*}
10: $f_{\mathscr{M}2} \leftarrow \text{trainModel}(\mathscr{A}, \sigma, \mathbf{m}, \mathbf{C})$ {*model retrain*}
11: $\mathbf{y} \leftarrow f_{\mathscr{M}2}([\mathbf{x}_1, \ldots, \mathbf{x}_\lambda])$ {2^{nd} *model prediction*}
12: $(\mathbf{y})_i \leftarrow y_{\text{orig},i}$ for all original-evaluated $y_{\text{orig},i} \in \mathbf{y}_{\text{orig}}$
13: $\alpha \leftarrow \text{selfAdaptation}(\mathbf{y}, \hat{\mathbf{y}})$
14: $\sigma, \mathbf{m}, \mathbf{C} \leftarrow$ CMA-ES update
15: **end while**
16: $\mathbf{x}_{\text{res}} \leftarrow \mathbf{x}_k$ from $\mathscr{A}$ where y_k is minimal

Output: $\mathbf{x}_{\text{res}}$ (point with minimal y)

2.2 aDTS-CMA-ES

The aDTS-CMA-ES [3, 19, 21], utilizes a GP surrogate model to estimate the fitness of a fraction of the population. A pseudocode is given in Algorithm 1. The algorithm expects an uncertainty criterion $\mathscr{C}$ for choosing solutions for re-evaluation. In optimization based on Gaussian processes, such criteria are conveniently defined on the marginal GP posterior, which is a univariate Gaussian distribution. One of the most prominent uncertainty criteria is the probability of improvement, $\mathscr{C}_{\text{POI}}(\mathbf{x}; T) = Pr(f(\mathbf{x}) \leq T)$, i. e., the posterior probability that the function value at a candidate solution $\mathbf{x}$ improves on a chosen target T, typically set to the historically best fitness value.

The sampling in aDTS-CMA-ES is identical to that of CMA-ES. The surrogate model is trained twice per generation. The first model is trained on a data set, which naturally cannot contain any individuals from the current population. A fraction α of the population is selected according to $\mathscr{C}$, evaluated with the (expensive) fitness function and included into the archive of individuals with known fitness values. The model is retrained and used to predict the remainder of the population. The fraction α is adapted according to surrogate model performance.

3 Gaussian Processes

Let $\mathscr{X}$ be some input space of dimensionality D. Gaussian process with a mean function $\mu\colon \mathscr{X} \to \mathbb{R}$ and a covariance function $k\colon \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, is a collection of random variables $(f(\mathbf{x}))_{\mathbf{x}\in\mathscr{X}}$ such that every finite-variate marginal $(f(\mathbf{x}_i))_{i=1}^N$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu(X), K(X,X))$, where $\mu(X) = (\mu(\mathbf{x}_i))_{i=1}^N$ and

$K(X,X) = (k(\mathbf{x}_i,\mathbf{x}_j))_{i,j=1}^{N}$. Both μ and k are parameterized, but we omit their parameters for the sake of readability.

3.1 Inference

Let $\mathbf{y} = \{y_1,\ldots,y_N\}$ be N $i.i.d.$ observations at inputs $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. A model with Gaussian likelihood and GP prior is given by distributions $\mathbf{y}\,|\,\mathbf{f} \sim \mathcal{N}(\mathbf{f},\sigma_n^2 I_N)$ and $\mathbf{f}\,|\,X \sim \mathcal{N}(\mu(X),K(X,X))$. From now on, we assume $\mu = 0$. Deterministic non-zero mean functions can be used by simply substracting from $\mathbf{y}$ (see [22] for more on this). Let us denote by θ the vector of hyperparameters consisting of parameters of k and noise variance σ_n^2.

The marginal likelihood of hyperparameters θ is (see [22])

$$p(\mathbf{y}\,|\,X,\theta) = \int p(\mathbf{y}\,|\,X,f,\theta)p(f\,|\,\theta)df \tag{1}$$

$$= \varphi(\mathbf{y}\,|\,\mathbf{0},K(X,X)+\sigma_n I_N), \tag{2}$$

where φ denotes the normalized multivariate Gaussian density.

In the regression problem, we are interested in conditional distribution $\mathbf{f}_*\,|\,\mathbf{y},X,X_*,\theta$ for X_* a set of N_* test inputs. Since $[\mathbf{y}^T\mathbf{f}_*^T]^T\,|\,X,X_*,\theta$ follows a multivariate Gaussian distribution, the distribution of $\mathbf{f}_*\,|\,\mathbf{y},X,X_*,\theta$ is also a multivariate Gaussian, in particular

$$\mathbf{f}_* \sim \mathcal{N}(\hat{\mathbf{f}}_*,\mathrm{cov}(\mathbf{f}_*)), \text{ where} \tag{3}$$

$$\hat{\mathbf{f}}_* = K(X_*,X)[K(X,X)+\sigma_n^2 I_N]^{-1}\mathbf{y} \tag{4}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n I_N]^{-1}K(X,X_*) \tag{5}$$

3.2 Hierarchical Model

When the covariance function family is given, model selection for GP regression is usually performed by maximum marginal likelihood estimate $\hat{\theta}_{\mathrm{ML}} = \arg\max_\theta \log p(y\,|\,X,\theta)$, which is a non-convex optimization problem. Computation of log marginal likelihood takes $\mathcal{O}(N^3)$ time due to a Cholesky decomposition of covariance matrix $K(X,X)$.

From a Bayesian perspective, especially if the number of hyperparameters is large or if N is small, it might be more appropriate to do inference with the marginal posterior distribution of hyperparameters

$$p(\theta\,|\,X,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,X,\theta)p(\theta)}{p(\mathbf{y}\,|\,X)}, \tag{6}$$

where $p(\mathbf{y}\,|\,X,\theta)$ is the marginal likelihood (1), now playing the role of the likelihood, and $p(\theta)$ is a hyper-prior. Simulations from $p(\theta\,|\,X,\mathbf{y})$ can be obtained by Bayesian computation methods, such as Markov chain Monte Carlo.

Uncertainty criteria in Algorithm 1 can thus incorporate uncertainty of hyperparameter estimation in addition to uncertainty about functions. In the current stage of research, we compute the prediction conditioned on a Bayes estimate $\theta_{\mathrm{Bayes}} = \mathrm{median}(\{\theta_s, s = 1,\ldots,S\})$, i. e., the median of the posterior sample.

4 Model Selection

If the probability of the true fitness function under GP prior is low, the performance of the model will be poor. For example, a GP with a neural network covariance fits data from a jump function better compared to a GP with a squared exponential [22] (more on covariance functions in Subsection 4.1). Searching over GP models with different covariances thus can be viewed as an automated construction of suitable priors. We select the model from a finite set according to a criterion of predictive performance, since this approach can easily be embedded into a combinatorial search algorithm, such as in [5]. GPs can represent random functions. The finite population of models included in our approach is described in Subsection 4.1. Some important classes of functions, such as linear and quadratic functions, neural networks and additive functions, are represented.

4.1 Model Population

The set of candidate GP models is shown in Figure 1. All models have zero mean.

A covariance function $k(\mathbf{x},\mathbf{x}')$ is stationary if it is a function of a distance $\|\mathbf{x}-\mathbf{x}'\|$. The squared exponential (SE) [22] is a stationary covariance function that leads to smooth processes [22].

A neural network (NN) covariance is a covariance of a GP induced by a Gaussian prior on weights of an infinitely wide neural network [18].

A dot product with a bias constant term models linear functions. The quadratic covariance is such a linear covariance squared. GPs with these covariances lead to Bayesian variants of linear and quadratic regression, respectively.

Additive covariance functions [6] are sums of lower dimensional components. We include an additive covariance function with a single degree of interaction – a superposition of one-dimensional squared exponentials.

Finally, we consider two cases of composite covariance functions: a sum of a squared exponential and a neural network; and a sum of a squared exponential and a quadratic.

4.2 Performance Criteria

We would like to select the surrogate model based on an estimation of out-of-sample predictive accuracy.

An attractive estimate of the out-of-sample predictive accuracy is cross-validation based on some partitioning of the data set into multiple data sets called folds. However, choosing among multiple GP models by cross-validation in each generation of the evolutionary optimization can

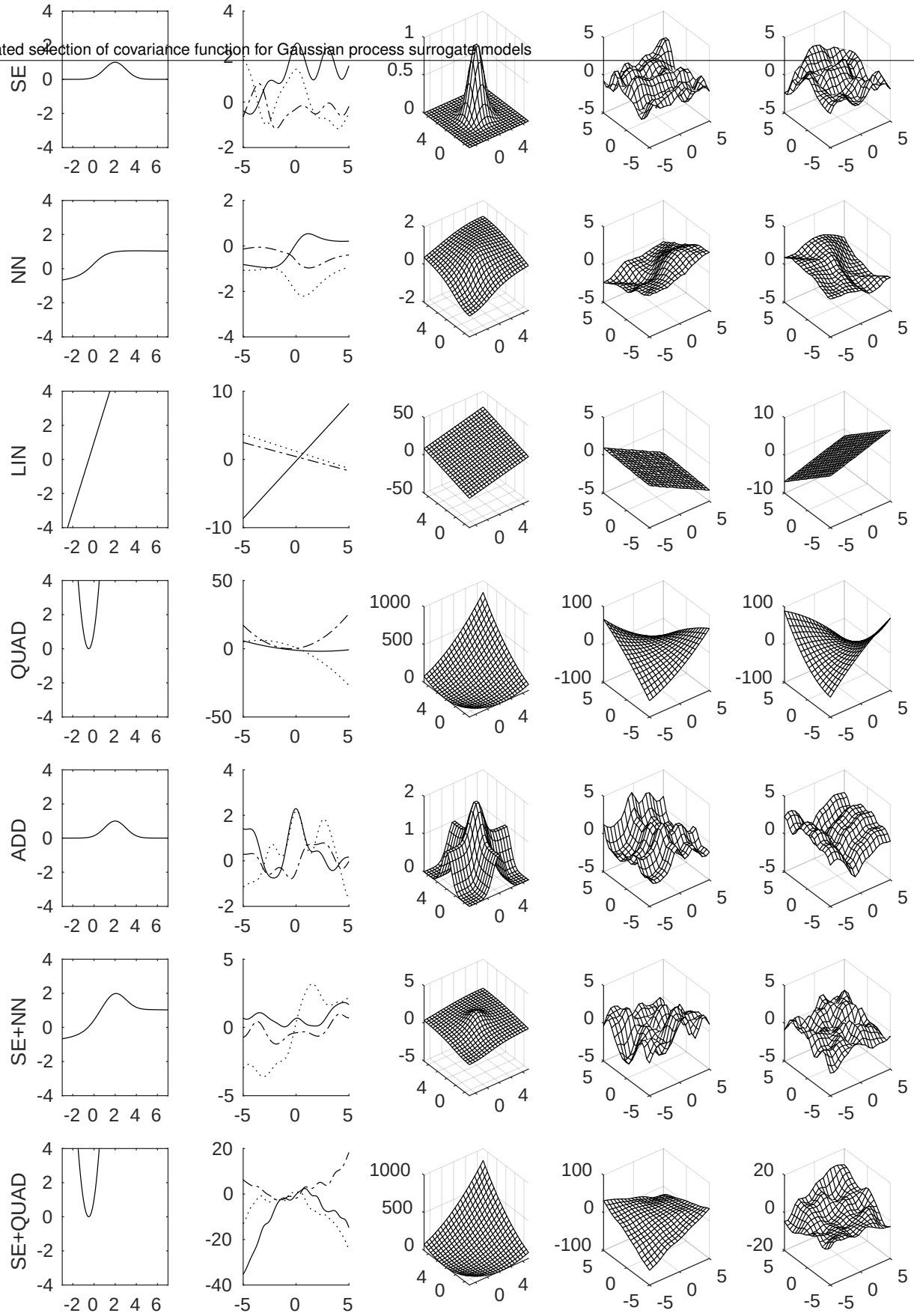

Figure 1: Rows: Used covariance functions. Columns 1–2: The covariance function on $\mathbb{R}$ centered at point 2 (Col. 1) and three independent samples from the GP (Col. 2). Columns 3–5: The covariance function on $\mathbb{R}^2$ centered at $[2\ 2]^T$ (Col. 3) and two independent samples from the GP (Col. 4 and 5).

be considered prohibitive from the computational perspective.

In the remainder of this subsection we follow the exposition of model comparison from Bayesian perspective given in [8]. We denote by q the true distribution from which data $\mathbf{y}$ are sampled and we suppress conditioning on X for simplicity.

A general measure of fit of a probabilistic model $\mathbf{y}$ to data is the log likelihood or log predictive density $\log p(\mathbf{y}\,|\,\theta) = \log\prod_{i=1}^{N} p(y_i\,|\,\theta)$. The quantity $-2\log p(y\,|\,\theta)$ is called deviance.

Akaike information criterion (AIC) [1] and related Bayes information criterion (BIC) [23] are based on the expected log predictive density conditioned on a maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}}$,

$$\mathrm{elpd}_{\hat{\theta}} = E_q(\log p(\tilde{\mathbf{y}}\,|\,\hat{\theta}_{\mathrm{ML}})), \tag{7}$$

where the expectation is taken over all possible data sets $\tilde{\mathbf{y}}$. Since expectation (7) cannot be computed exactly, it is estimated from sample $\mathbf{y}$. The AIC and BIC compensate for the bias towards overfitting by substracting a correction term, the number of parameters n_θ and $\frac{1}{2}n_\theta \log N$, respectively.

For hierarchical Bayesian models, such as (6), it is not always entirely clear, what the parameters of the model are, since the likelihood can factorize in different ways. The deviance information criterion (DIC) [24] is still based on deviance, conditioned on a Bayes estimate $\hat{\theta}_{\mathrm{Bayes}}$, but the effective number of parameters p_{DIC} depends on data. We define the DIC for the marginal likelihood (1), focusing on hyperparameters θ, although it could be defined for the likelihood $p(\mathbf{y}\,|\,f,\theta)$, focusing on both f and θ.

We use the following definition of the effective number of parameters (see [8]):

$$p_{\mathrm{DIC}} = 2\mathrm{var}_{\mathrm{post}}(\log p(\mathbf{y}\,|\,\theta)),$$

which can be estimated by the sample variance of a posterior sample. Using the effective number of parameters, the DIC is

$$DIC = -2\log p(\mathbf{y}\,|\,\hat{\theta}_{\mathrm{Bayes}}) + 2p_{\mathrm{DIC}}.$$

A probabilistic model is called regular if its parameters are identifiable and its Fisher information matrix is positive definite for all parameter values. The model is called singular otherwise. The information criteria defined above assume regularity. The Widely applicable information (WAIC) [26] works also for singular models. The WAIC is based on estimation of the expected log *pointwise* predictive density

$$\begin{aligned}\mathrm{elppd} &= \sum_{i=1}^{N} E_q(\log p_{\mathrm{post}}(\tilde{y}_i)) \\ &= \sum_{i=1}^{N} E_q(\log \int p(\tilde{y}_i\,|\,\mathbf{y},\theta)p(\theta\,|\,\mathbf{y})d\theta.\end{aligned}$$

The estimation of elppd from the sample is biased, so again, an effective number of parameters must be added as a correction. We use the following definition of the WAIC (see [8]):

$$WAIC = -\sum_{i=1}^{N}\log p_{\mathrm{post}}(y_i) + \sum_{i=1}^{N}\mathrm{var}_{\mathrm{post}}(\log p(y_i\,|\,\theta)),$$

that is the negative log pointwise predictive density corrected for bias by pointwise posterior variance of log predictive density.

The pointwise predictive density $p_{\mathrm{post}}(y_i\,|\,\mathbf{y},\theta)$ for the GP model (1) is computed by integrating Gaussian likelihood over the marginal posterior GP at i^{th} training point:

$$\begin{aligned}p(y_i\,|\,\mathbf{y},\theta) &= \int p(y_i\,|\,\mathbf{y},f_i,\theta)p(f_i\,|\,\mathbf{y},\theta)\,df_i \\ &= \varphi(y_i\,|\,\hat{f}_i,\sigma_n^2 + \mathrm{var}(f_i)),\end{aligned}$$

where φ denotes the Gaussian density and $\hat{f}_i, \mathrm{var}(f_i)$ are as in (3).

5 Experimental Evaluation

In this section, we describe preliminary experimental evaluation procedure of aDTS-CMA-ES that uses a GP model with an automated selection of covariance function. Since GPs are a nonparametric model, we opt for the WAIC, which require a sample from distribution (6). We use Metropolis-Hastings MCMC with an adaptive proposal distribution [9] [2].

Algorithm 1 is updated in the following way: [3]

1. In steps (5) and (10), all GPs from Figure 1 are trained.
2. The predictive accuracy of all models is evaluated using the WAIC (4.2). The DIC (4.2) is also computed for information, but not taken into account.
3. The model with the lowest WAIC is used for prediction (steps (6) and (11)).

The hyper-priors are chosen as follows: log-normal with mean $\log(0.01)$ and variance 2 for σ_n^2; and log-$t_{\nu=4}$ with mean 0 for all other hyperpameters.

5.1 Setup

The proposed algorithm implemented in MATLAB is evaluated on the noiseless testbed of the COCO/BBOB (Comparing Continuous Optimizers / Black-Box Optimization Benchmarking) framework [11,12] and compared with the GP-based aDTS-CMA-ES and the CMA-ES itself.

[2] Using MATLAB implementation available at `http://helios.fmi.fi/~lainema/dram/`

[3] The sources are available at `https://github.com/repjak/surrogate-cmaes/tree/modelsel`

The testbed consists of 24 functions, each defined everywhere on $\mathbb{R}^D$ with the optimum in $[-5,5]^D$ for all dimensionalities $D \geq 2$. Each test function has multiple instances which are derived by various transformations of input space or f-space. We run the algorithm on 5 instances $(1 \ldots, 5)$ as opposed to 15 recommended instances for the reason of increased computational demands of the modified algorithm. For the same reason, only functions of 10 variables ($10D$) are considered.

If not stated otherwise, all settings of the aDTS-CMA-ES are as recommended in [3].

The CMA-ES results in BBOB format were downloaded from the BBOB 2010 workshop archive [4].

5.2 Results

Figure 2 gives the scaled best-achieved logarithms $\Delta_f^{\log}$ of median distances to the functions optimum for the respective number of function evaluations per dimension (FE/D). Medians and the 1st and 3rd quartiles are calculated from 5 independent instances in case of the algorithm with covariance selection according to the WAIC and from 15 independent instances otherwise. We observe that in most cases, the WAIC-based algorithm mostly barely outperforms the pure CMA-ES, which suggests the chosen model is generally weak and the adaptivity mechanism basically turns off using the surrogate model. The functions where the WAIC variant outperforms the aDTS-CMA-ES (f21 and f22) are multi-modal and the interquartile range is large.

In order to compare the considered information criteria, we calculate the rank of each model under both WAIC and DIC. Table 1 summarizes the average ranks over all model selections performed on each benchmark function. We observe that the DIC often prefers the additive model, while the WAIC is more balanced in this respect. Surprisingly the linear kernel has been very rarely selected even on the linear function (f5) under both information criteria. A similar observation holds for the quadratic kernel and the quadratic functions (f1, f2).

6 Conclusion & Further Work

In this paper, we presented an algorithm for selecting a GP kernel using Bayesian model comparison techniques. Preliminary experiments for the model selection plugged into the aDTS-CMA-ES algorithm were conducted on the COCO/BBOB testbed. Due to the small number of experiments performed so far, it is difficult to draw any serious conclusions. The first obtained results may indicate improper convergence of the MCMC sampler or that more sophisticated covariance functions may be needed.

One direction of future research, beside analyzing and repairing aforementioned deficiencies, is an extension of the proposed algorithm into a combinatorial search over kernels in flavor of [5,7], which is challenging due to computational costs related to the need of repeated surrogate model retraining.

One possible direction of research is a *co-evolution* of a population of covariance functions alongside the population of candidate solutions to the black-box objective function. Other related research area is applying surrogate modeling to high-dimensional problems using algorithms for variable selection via multiple kernel learning [2, 6].

[4] `http://coco.gforge.inria.fr/data-archive/bbob/2010/`

Acknowledgements

This research was supported by SVV project number 260 453 and the Czech Science Foundation grants No. 17-01251. Further, access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

[1] H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, 1973.

[2] F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning, 2009.

[3] L. Bajer. *Model-based evolutionary optimization methods*. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, 2018.

[4] L. Bajer, Z. Pitra, and M. Holeňa. Benchmarking Gaussian processes and random forests surrogate models on the BBOB noiseless testbed. In *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference - GECCO Companion '15*. Association for Computing Machinery (ACM), 2015.

[5] D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1166–1174, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[6] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen. Additive gaussian processes. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 226–234. Curran Associates, Inc., 2011.

[7] C. Gagné, M. Schoenauer, M. Sebag, and M. Tomassini. Genetic programming for kernel-based learning with co-evolving subsets selection. In *PARALLEL PROBLEM SOLVING FROM NATURE, REYKJAVIK, LNCS*, pages 1008–1017. Springer Verlag, 2006.

[8] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, Nov. 2014.

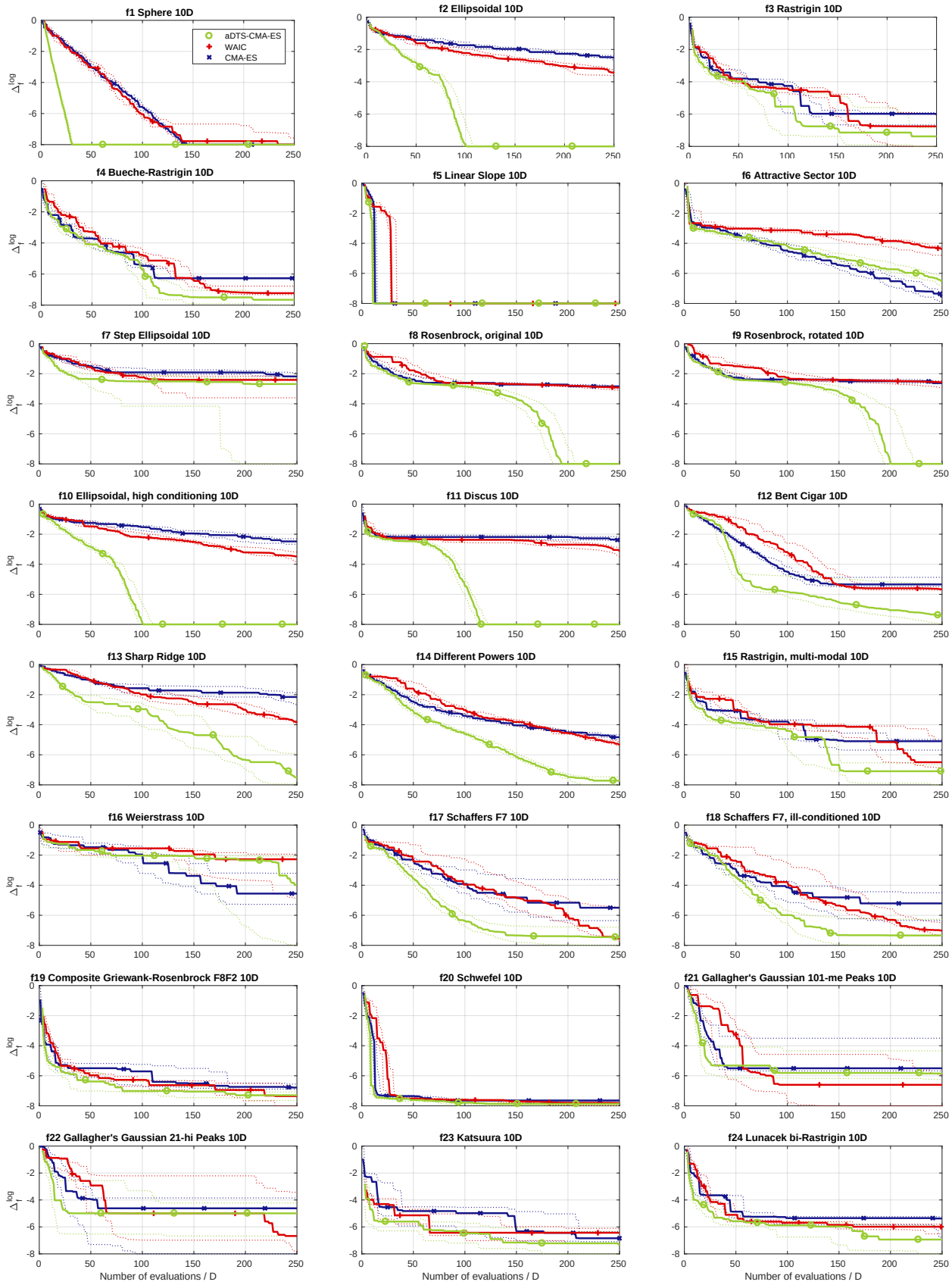

Figure 2: Medians (solid) and 1st/3rd quartiles (dotted) of the distances to the optima of 24 COCO/BBOB benchmarks in 10D for algorithms aDTS-CMA-ES (green), aDTS-CMA-ES with WAIC-based model selection (red) and CMA-ES (blue). The medians and quartiles for WAIC variant were calculated from 5 independent instances. In all other cases, 15 indepenent instances were used. Distances to optima are shown in the $\log_{10}$ scale.

Table 1: Average model ranks in 10D for each predictive performance criterion. The lowest value in bold.

Criterion	WAIC							DIC						
Model	SE	NN	LIN	QUAD	ADD	SE+NN	SE+LIN	SE	NN	LIN	QUAD	ADD	SE+NN	SE+LIN
f1	3.64	4.01	6.94	4.17	**2.73**	3.25	3.27	5.57	4.68	6.65	3.11	**1.55**	4.07	2.37
f2	3.07	3.05	6.96	5.41	4.35	**2.48**	2.67	5.38	4.85	6.78	3.99	**1.45**	3.71	1.84
f3	2.94	3.20	6.75	5.99	4.14	**2.45**	2.53	4.37	4.63	6.74	5.40	**1.40**	3.11	2.36
f4	3.00	3.20	6.87	5.73	4.11	**2.52**	2.57	4.49	4.67	6.76	5.21	**1.30**	3.27	2.30
f5	3.69	3.60	6.78	4.41	**2.69**	3.28	3.56	6.75	4.16	4.71	3.25	3.48	3.00	**2.66**
f6	3.03	3.09	6.99	5.95	3.99	**2.46**	2.50	4.18	4.62	6.92	5.92	1.86	2.80	**1.69**
f7	3.15	3.09	6.92	5.86	3.91	**2.49**	2.58	4.35	4.87	6.90	5.37	**1.54**	2.99	1.98
f8	2.83	3.12	6.97	5.76	4.15	**2.50**	2.66	4.78	4.57	6.83	5.21	**1.32**	3.27	2.02
f9	2.86	3.14	6.97	5.69	4.14	2.63	**2.57**	4.86	4.57	6.79	5.05	**1.32**	3.37	2.04
f10	3.00	3.16	6.96	5.43	4.31	**2.52**	2.62	5.39	4.82	6.76	3.95	**1.53**	3.79	1.75
f11	3.01	3.09	6.97	5.53	4.24	**2.57**	2.60	5.21	5.17	6.80	3.66	2.14	3.87	**1.16**
f12	3.16	3.28	6.93	4.37	4.96	**2.64**	2.66	5.46	4.65	6.66	3.57	**1.46**	3.94	2.27
f13	2.93	2.98	6.98	5.83	4.42	**2.37**	2.50	4.87	4.47	6.88	5.23	**1.24**	3.06	2.25
f14	3.29	2.96	7.00	5.90	3.63	**2.59**	2.64	4.26	4.84	6.99	5.64	**1.43**	3.04	1.81
f15	2.86	3.28	6.73	5.95	4.13	2.58	**2.47**	4.25	4.69	6.80	5.58	**1.48**	2.92	2.28
f16	2.53	3.59	6.46	6.50	4.16	2.41	**2.36**	3.69	4.69	6.78	6.10	**1.80**	2.44	2.51
f17	3.25	3.05	6.86	6.09	3.52	**2.58**	2.65	4.03	4.69	6.90	5.95	**1.36**	2.80	2.26
f18	3.17	3.05	6.88	6.10	3.65	**2.55**	2.60	3.98	4.70	6.88	6.00	**1.42**	2.74	2.28
f19	2.60	3.52	6.42	6.54	4.20	**2.32**	2.39	3.69	4.67	6.75	6.13	**1.45**	2.52	2.78
f20	2.98	3.37	6.94	5.55	4.01	2.62	**2.51**	4.46	4.73	6.81	5.22	**1.30**	3.26	2.21
f21	3.14	3.21	6.87	5.07	4.61	2.59	**2.51**	4.91	4.65	6.76	4.56	**1.34**	3.47	2.31
f22	3.11	3.22	6.88	5.00	4.57	**2.58**	2.63	5.06	4.60	6.73	4.31	**1.41**	3.51	2.37
f23	2.47	3.77	6.35	6.58	4.28	2.34	**2.21**	3.50	4.76	6.68	6.15	**1.74**	2.47	2.70
f24	2.73	3.53	6.44	6.51	4.06	2.38	**2.34**	3.76	4.68	6.77	6.11	**1.42**	2.53	2.73

[9] H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, Dec 2006.

[10] N. Hansen. *The CMA evolution strategy: a comparing review*, pages 75–102. Springer, Berlin, Heidelberg, 2006.

[11] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter Black-Box Optimization Benchmarking 2009: Noiseless functions definitions. Technical report, INRIA, 2009, updated 2010.

[12] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter Black-Box Optimization Benchmarking 2012: Experimental setup. Technical report, INRIA, 2012.

[13] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, June 2001.

[14] S. Kern, N. Hansen, and P. Koumoutsakos. *Local Metamodels for Optimization Using Evolution Strategies*, pages 939–948. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[15] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. *CoRR*, abs/1402.4304, Apr. 2014.

[16] I. Loshchilov, M. Schoenauer, and M. Sebag. Intensive surrogate model exploitation in self-adaptive surrogate-assisted cma-es (saacm-es). In *Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference - GECCO '13*. ACM Press, 2013.

[17] J. Močkus. On bayesian methods for seeking the extremum. In *Proceedings of the IFIP Technical Conference*, London, UK, 1974. Springer-Verlag.

[18] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[19] Z. Pitra, L. Bajer, and M. Holeňa. Doubly trained evolution control for the surrogate CMA-ES. In *Parallel Problem Solving from Nature – PPSN XIV*, pages 59–68. Springer International Publishing, 2016.

[20] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. Overview of surrogate-model versions of covariance matrix adaptation evolution strategy. In *Proceedings of the Genetic and Evolutionary Computation Conference 2017, Berlin, Germany, July 15–19, 2017 (GECCO '17)*. ACM, July 2017.

[21] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. Adaptive Doubly Trained Evolution Control for the Covariance Matrix Adaptation Evolution Strategy. In *ITAT 2017: Information Technologies—Applications and Theory*, Martin, Sept. 2017. CreateSpace Independent Publ. Platform.

[22] C. E. Rassmusen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning series. MIT Press, 2006.

[23] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[24] D. Spiegelhalter, N. Best, B. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):583–616, 12 2002.

[25] H. Ulmer, F. Streichert, and A. Zell. Evolution strategies assisted by Gaussian processes with improved pre-selection criterion. In *The 2003 Congress on Evolutionary Computation, 2003. CEC '03.*, pages 692–699. Institute of Electrical and Electronics Engineers (IEEE), 2003.

[26] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11:3571–3594, Dec. 2010.