Landscape Analysis of Surrogate-based Model Optimization Tasks

Patrick Koch

Cologne University of Applied Sciences
Campus Gummersbach

1 September 2012

Motivation

Optimization of supervised machine learning parameters (model optimization)

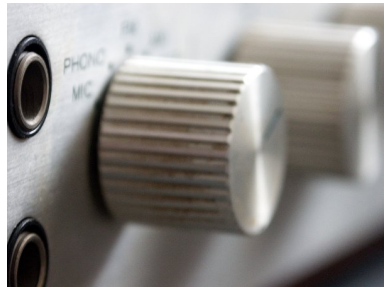
How to set the parameters of a machine learning (ML) process?

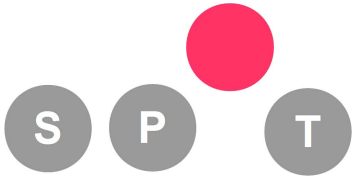
- Trial and error (hand-tuning, often done, but probably not the best solution)
- Local search (many existing heuristics, but might get stuck in local optima)
- Global optimization (can be expensive, seldom done)

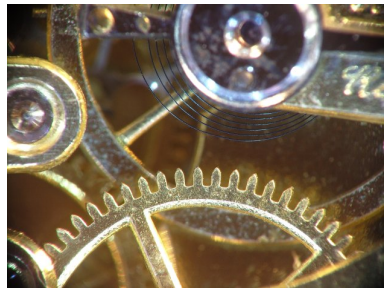
Tuned Data Mining in R

Konen *et al.* [Kon11] developed the Tuned Data Mining in R (TDMR)[1] software framework for easily setting up machine learning experiments:

- Methods for continuous optimization (supporting Box-constraints):
  - Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [HO01],
  - Direct search algorithms (Kolda *et al.* [KLT03])
  - ...
- Efficient global optimization (EGO) [JSW98]
  - Sequential Parameter Optimization (SPO) [BBLP05], combines classical design of experiments [Fis36] with design and analysis of computer experiments (DACE) [SWMW89]



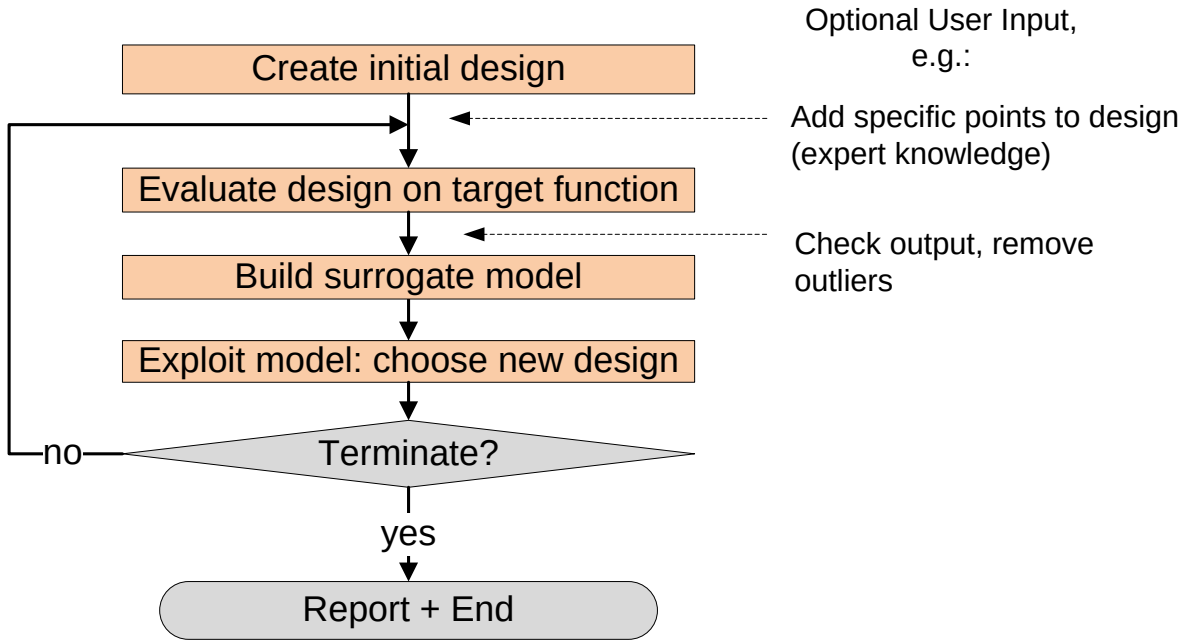




[1] `http://cran.r-project.org/web/packages/TDMR/index.html`

Efficient Global Optimization (EGO)

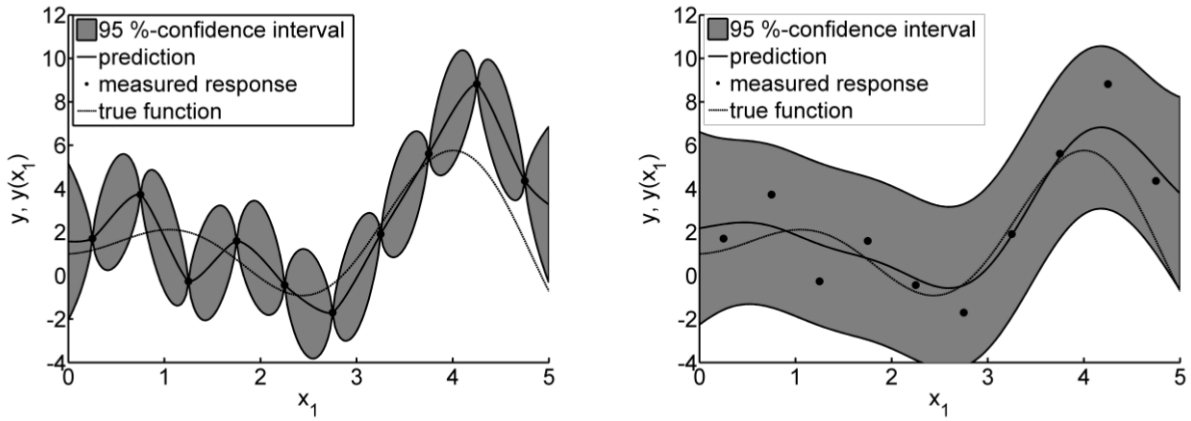
- EGO algorithms like SPO perform the main part of the optimization on the cheap surrogate function and evaluate only some points on the expensive real function to refine the surrogate:



- Konen *et al.* [KKF+11] compared optimization techniques for ML, where SPO with Kriging performed best

Surrogate Models: Kriging

- Kriging was developed by Matheron [Mat63] and named after the mining engineer D.G. Krige [Kri51]
- The idea is to model the responses $y(\vec{x})$ creating a surrogate function $\hat{y}(\vec{x})$ based on Gaussian processes.
- We differ between interpolating (left) and non-interpolating (right) DACE models with nugget term:



- The nugget term avoids the exact interpolation of the observations. The influence of outliers is relaxed and a smoother model can be computed [Wag10].

Landscape Analysis

In the benchmark by Konen *et al.* [KKF+11] SPO with Kriging performed best, but how good are the Kriging landscapes actually?

Q-1 How exact are Kriging metamodels?

Q-2 Are the landscapes accurate in all regions of the search space?

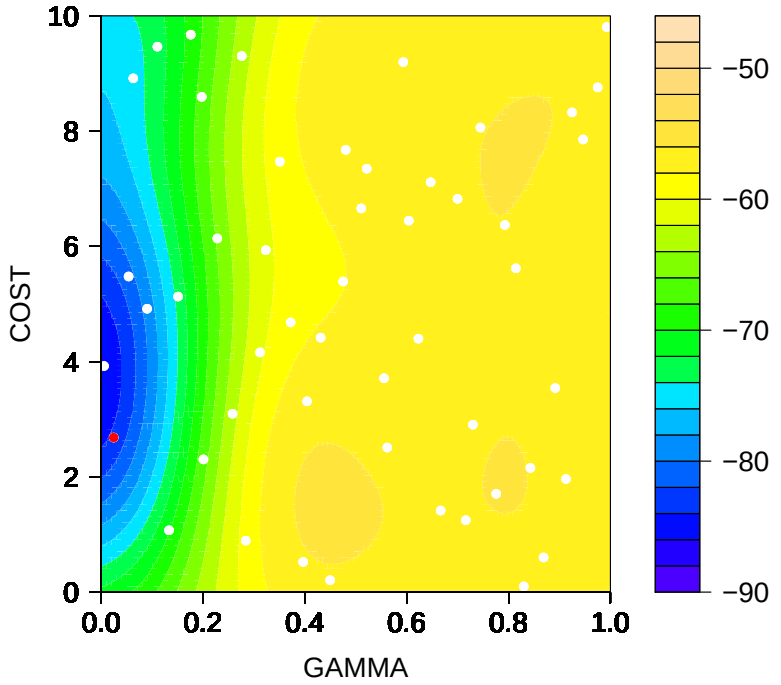
Q-3 Which effect has the choice of the Kriging method (interpolating vs. non-interpolating Kriging)?

$\Rightarrow$ We analyze the underlying fitness landscapes obtained during the optimization

The main challenge: considerable amount of *noise* in objective function for most supervised machine learning tasks

Landscape Analysis

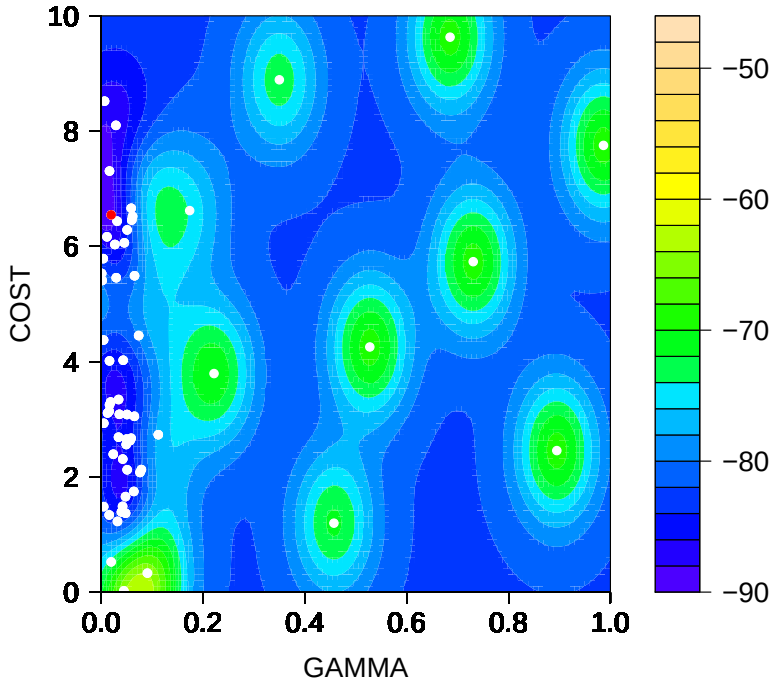
Kriging landscapes of interpolating Kriging metamodels without nugget estimation:



- Latin hypercube design (LHD) with interpolating Kriging (no nugget effect)
- SPO and interpolating Kriging (no nugget effect)

Landscape Analysis

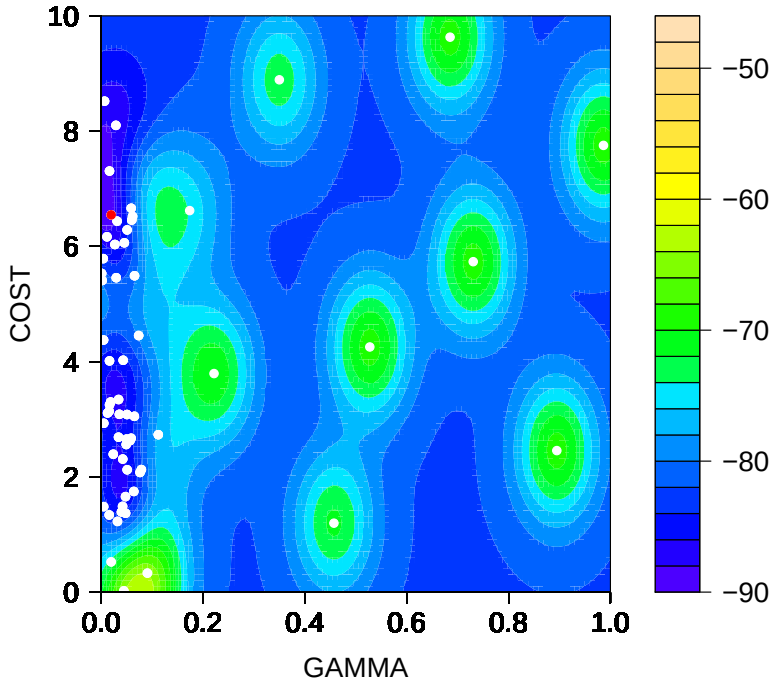
Kriging landscapes of interpolating Kriging metamodels without nugget estimation:



- SPO and interpolating Kriging (no nugget effect)
- Depending on the distribution of the design points, interpolating Kriging models can be misleading in certain regions! (Q-1, Q-2)

Landscape Analysis

Kriging landscapes of interpolating Kriging metamodels without nugget estimation:



- Latin hypercube design (LHD) with interpolating Kriging (no nugget effect)

- Depending on the distribution of the design points, interpolating Kriging models can be misleading in certain regions! (Q-1, Q-2)

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

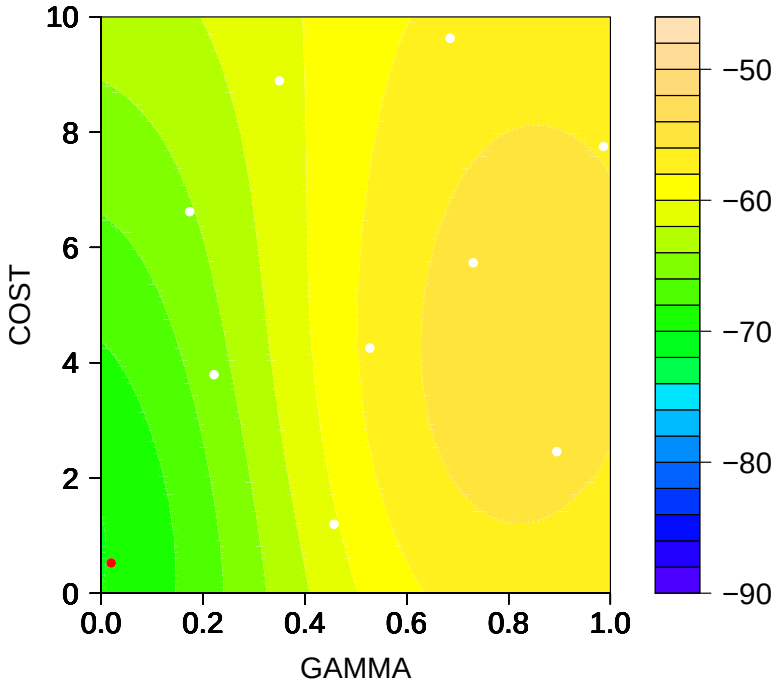


Figure: Estimated landscape after 10 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

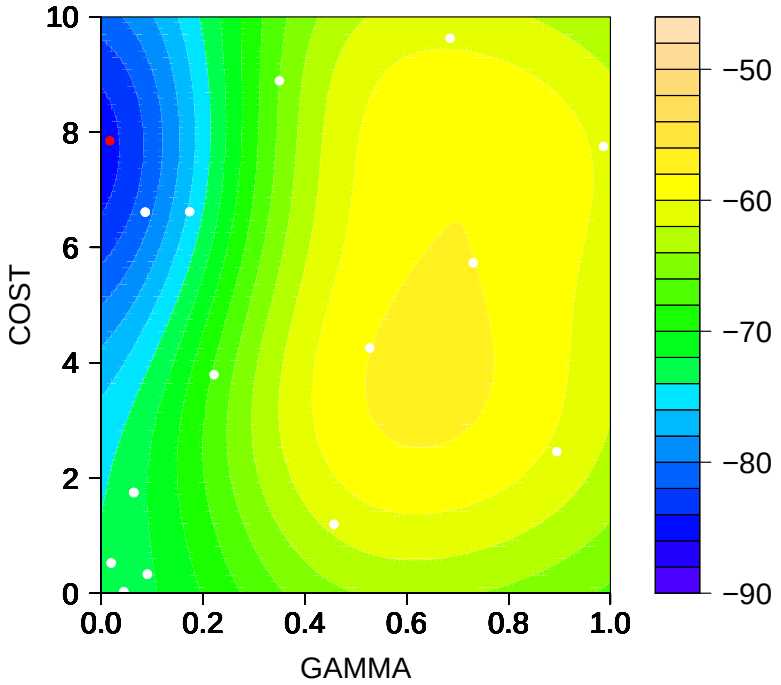


Figure: Estimated landscape after 20 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

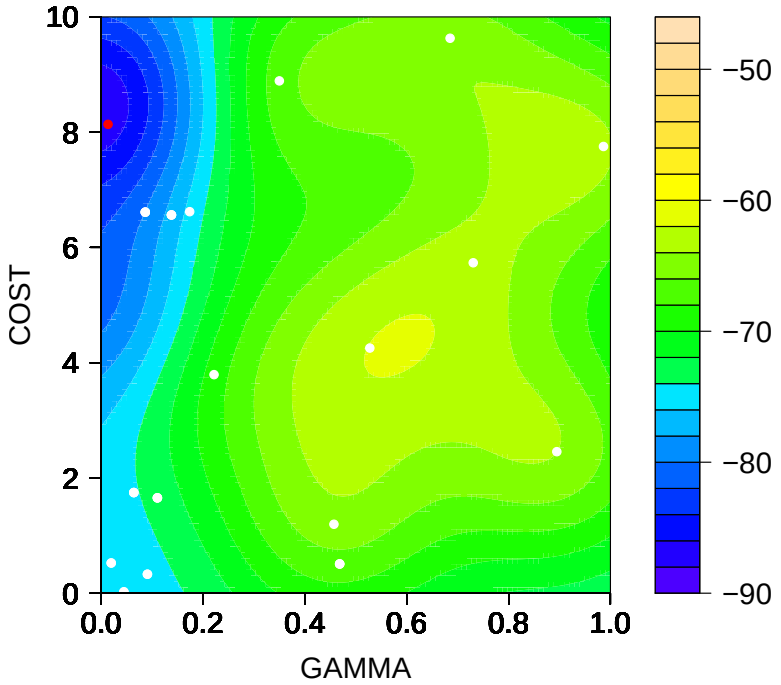


Figure: Estimated landscape after 30 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

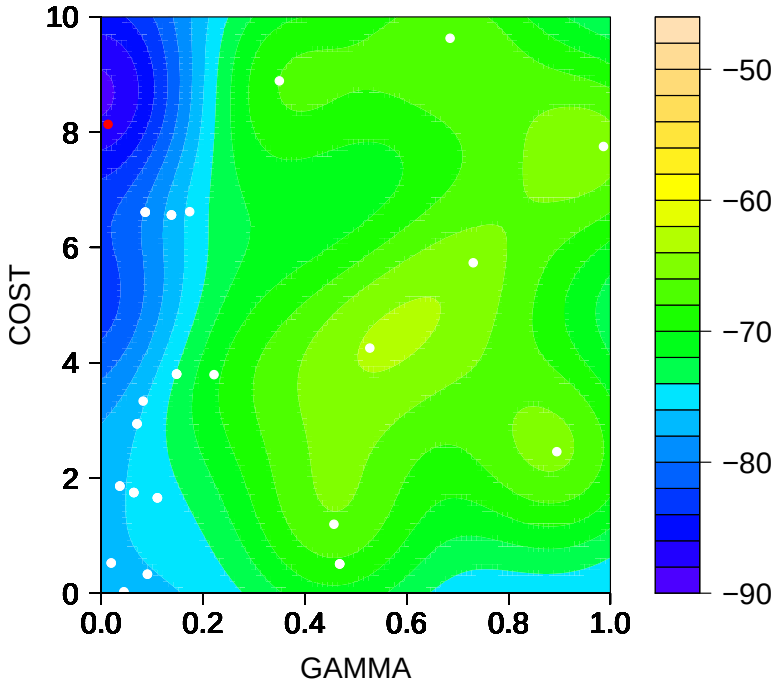


Figure: Estimated landscape after 40 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

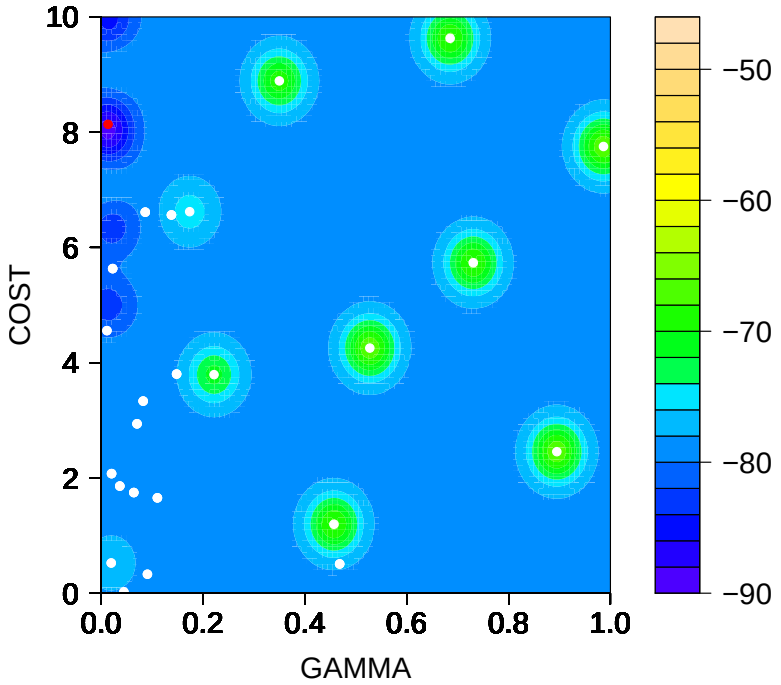


Figure: Estimated landscape after 50 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

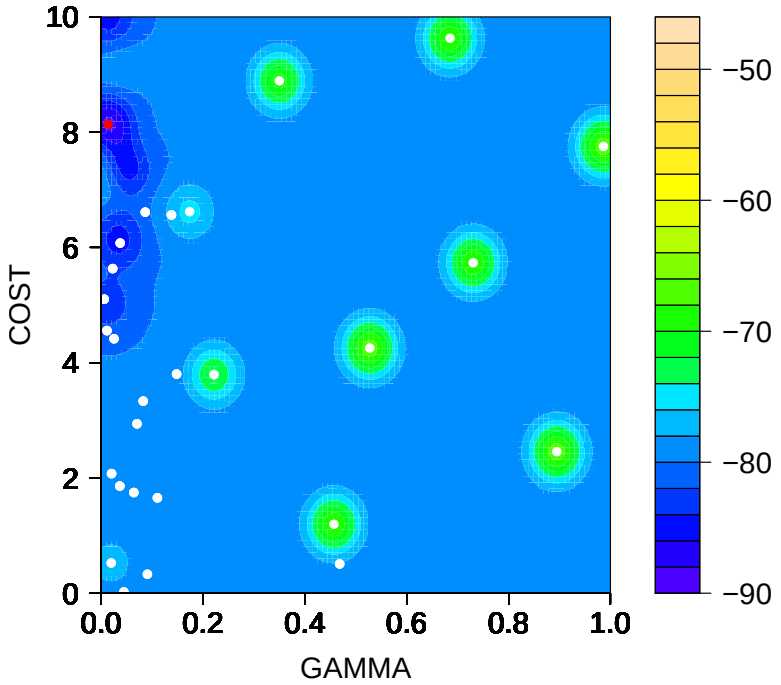


Figure: Estimated landscape after 60 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

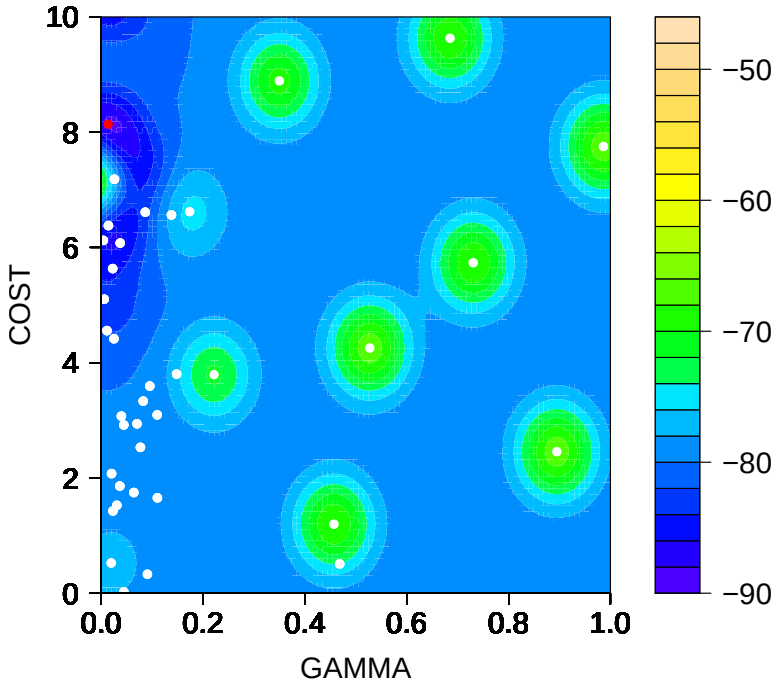


Figure: Estimated landscape after 90 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

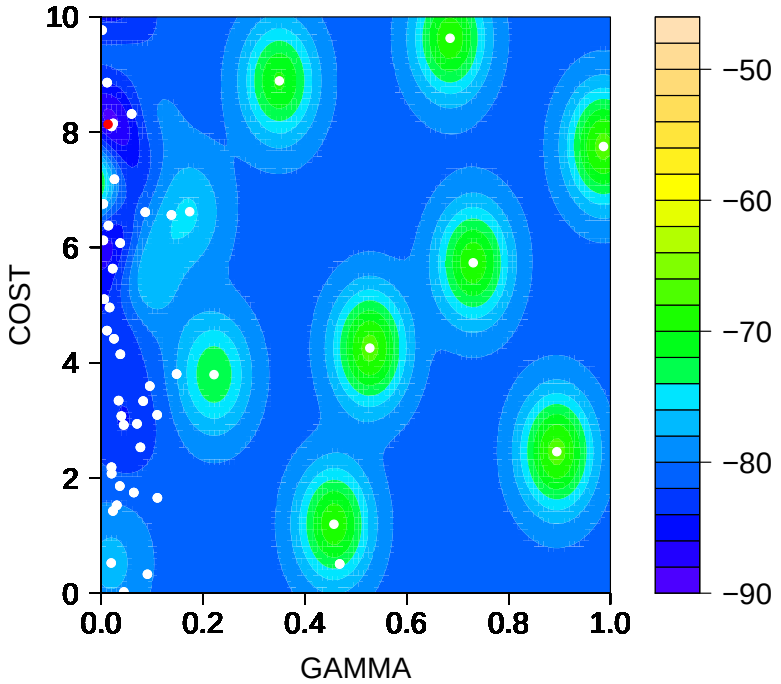


Figure: Estimated landscape after 120 objective function evaluations.

Interpolating Kriging metamodels

We show how additional design points deteriorate the fit:

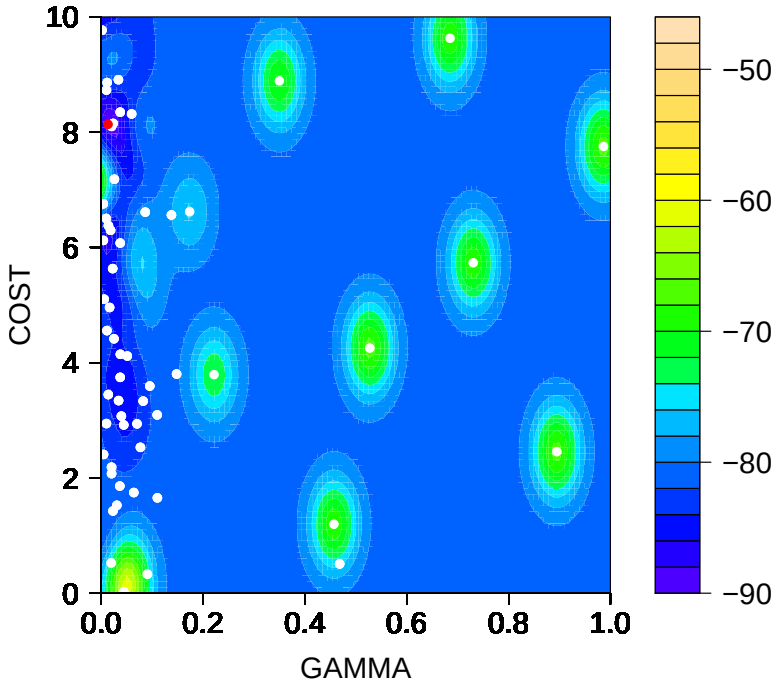


Figure: Estimated landscape after 150 objective function evaluations.

How (sometimes) wrong minima are detected

The greedy infill strategy of SPO together with a noisy objective function (small correlation lengths) and a too small initial design can lead to premature convergence of the search and deceptive landscapes:

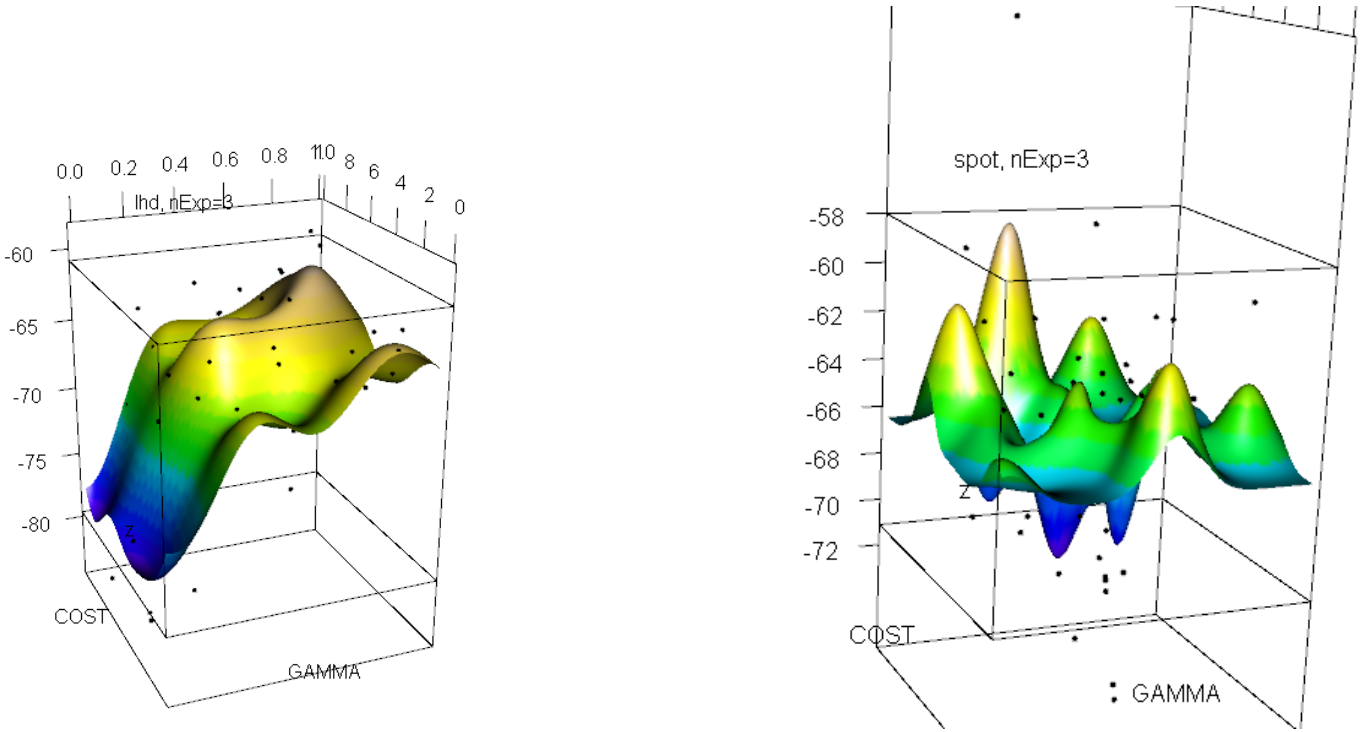


Figure: More realistic LHD (left) and SPO-approximated (right) landscapes for the same problem (ionosphere) with interpolating Kriging model.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

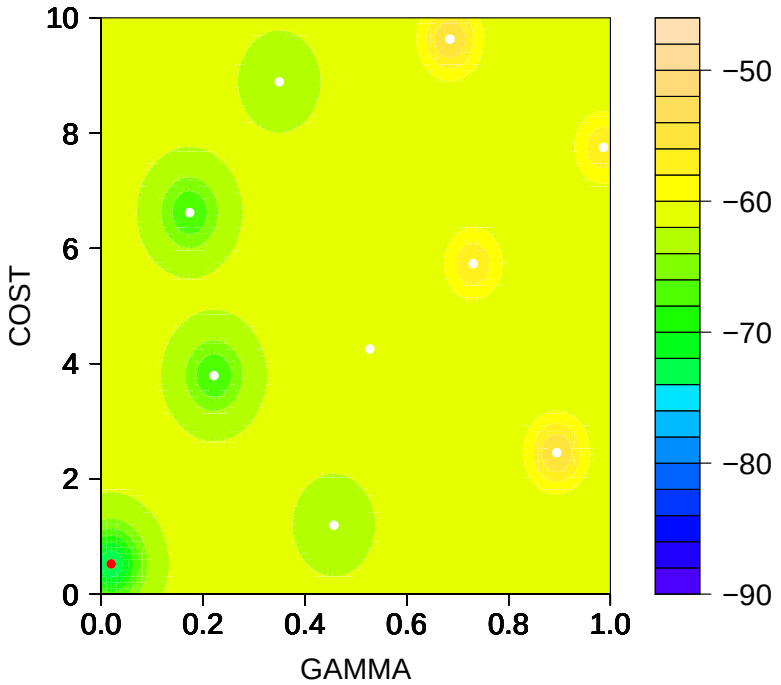


Figure: Estimated landscape after 10 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

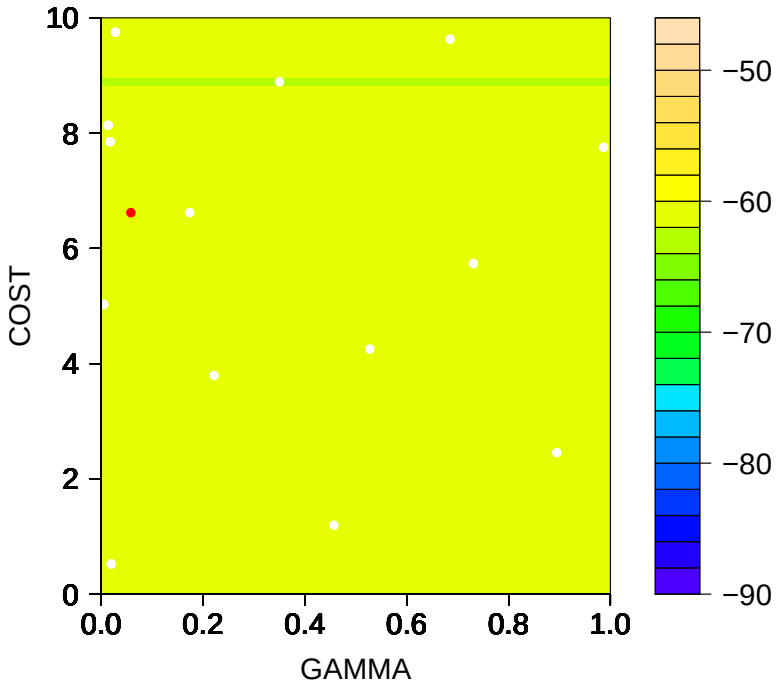


Figure: Estimated landscape after 20 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

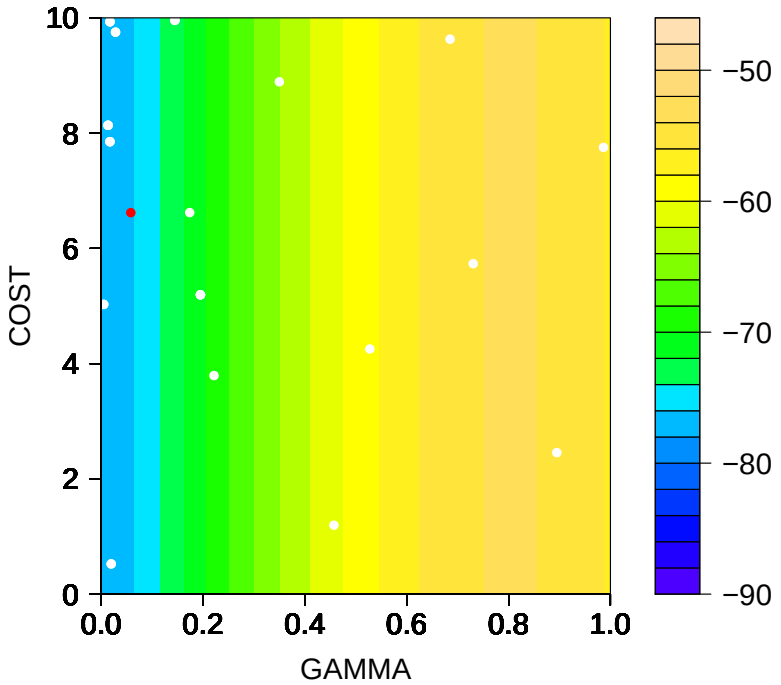


Figure: Estimated landscape after 30 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

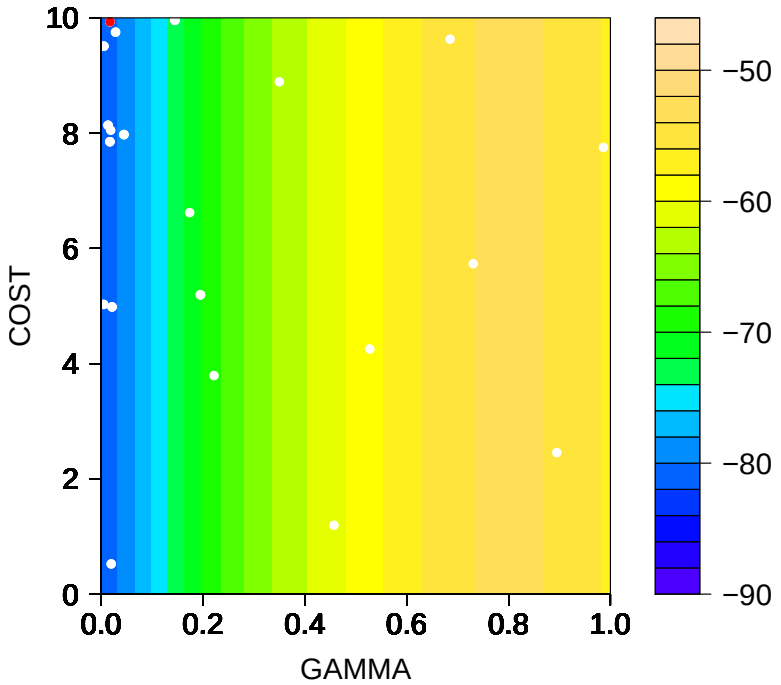


Figure: Estimated landscape after 40 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

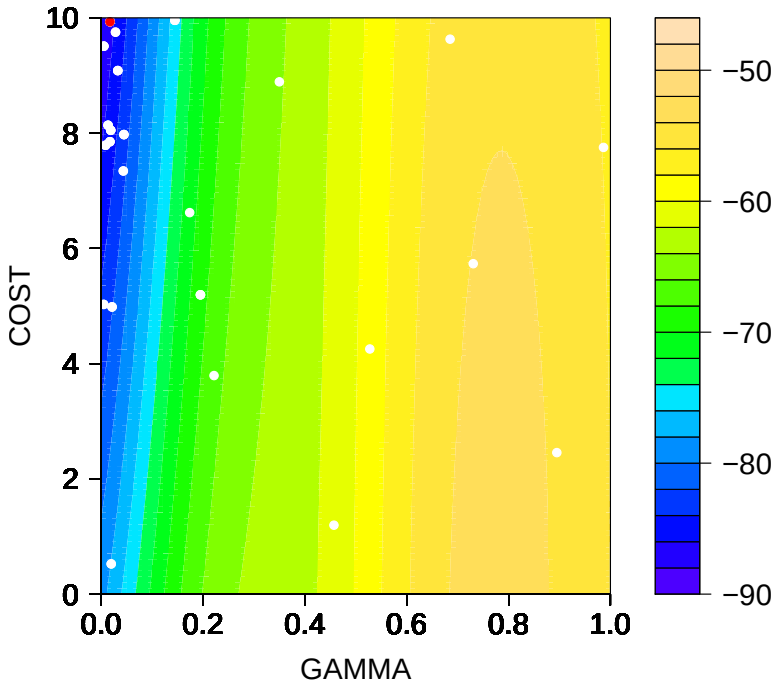


Figure: Estimated landscape after 50 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

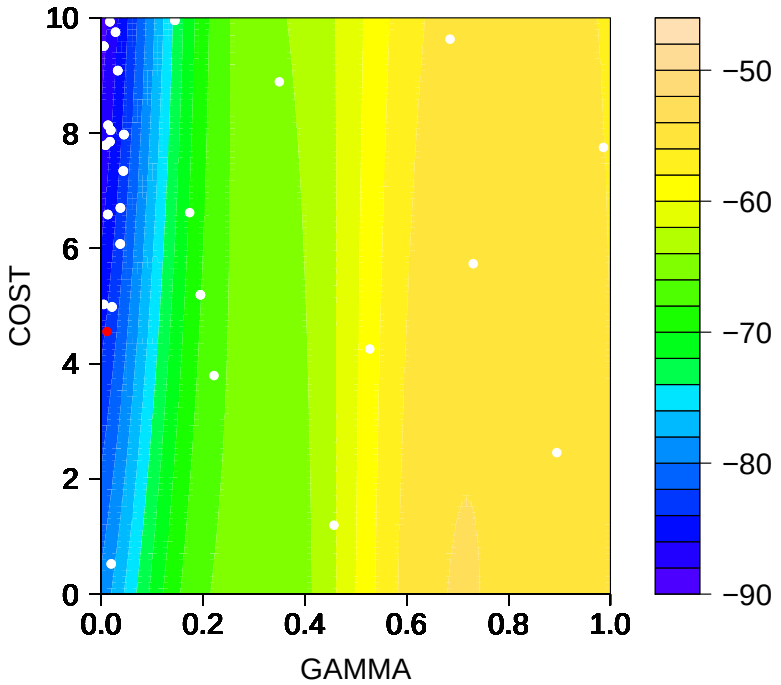


Figure: Estimated landscape after 60 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

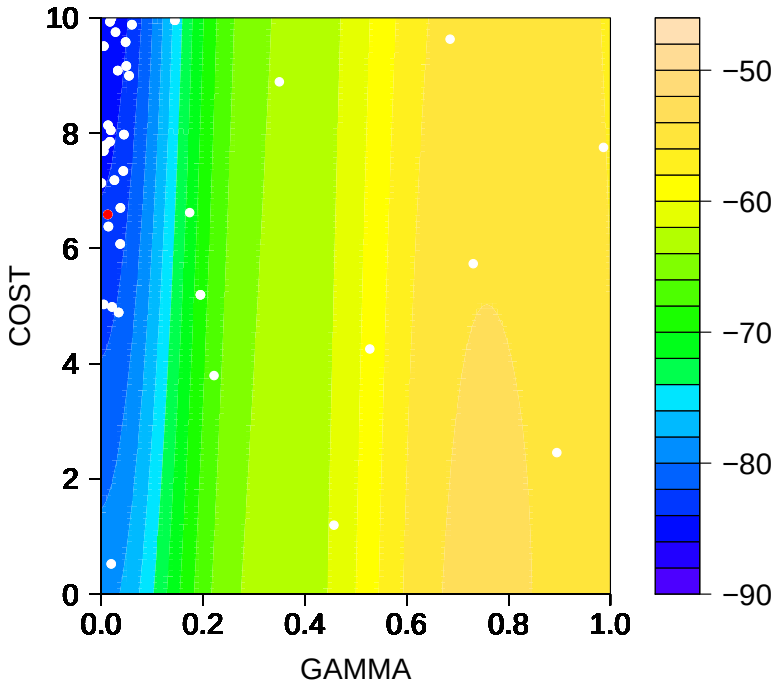


Figure: Estimated landscape after 90 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

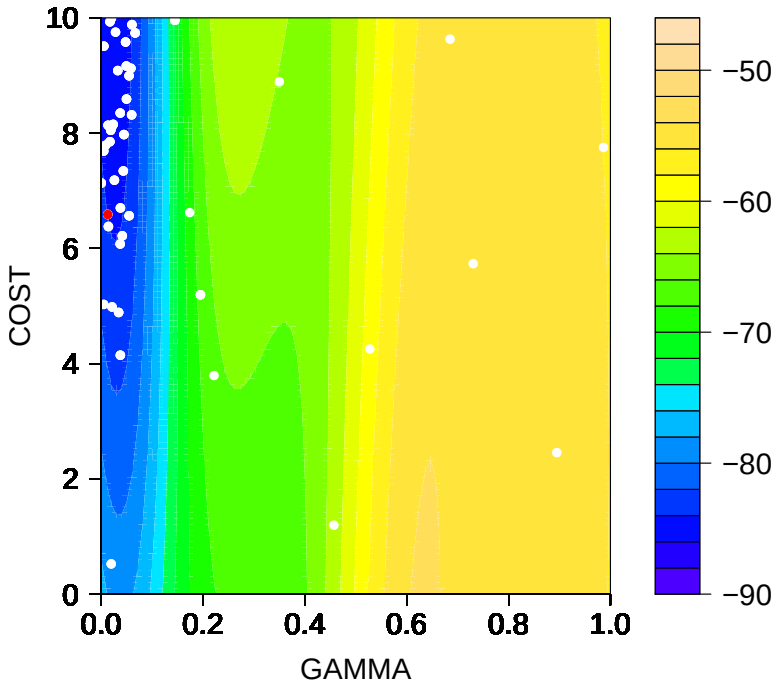


Figure: Estimated landscape after 120 objective function evaluations.

Non-interpolating Kriging metamodels

Now, using a non-interpolating Kriging method with nugget effect we get a much better estimate than with standard interpolating Kriging (Q-3):

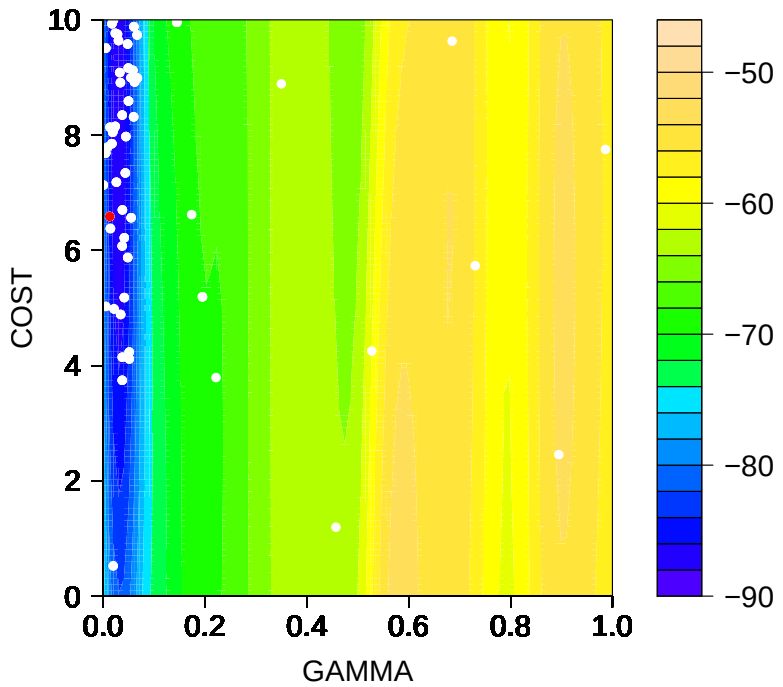


Figure: Estimated landscape after 150 objective function evaluations.

Exploration vs. Exploitation

- In our experiments a very greedy infill strategy was used to generate new points
- More exploration is necessary, respecting the uncertainties of the model
- State-of-the-Art is to maximize the Expected Improvement (EI) [SWJ98]:

$$EI(\vec{x}) = (y^* - \hat{y}(\vec{x}))\Phi(u(\vec{x})) + \hat{s}(\vec{x})\Phi(u(\vec{x})), u(\vec{x}) = \frac{y^* - \hat{y}(\vec{x})}{\hat{s}(\vec{x})} \quad (1)$$

- Unfortunately, non-interpolating Kriging models can't use this directly!

Expected Improvement for Noisy Observations

Forrester *et al.* [FKB06] proposed a technique called Re-interpolating Kriging, making it possible to use EI together with noisy observations:

1. Build a non-interpolating Kriging model based on the noisy observations
2. Create a large design of new design points
3. Apply the model to these points
4. Use the predictions of (3) to compute a new interpolating Kriging model
5. Maximize EI as usual using the interpolating Kriging model.

Summary

- We performed a parameter tuning for ML parameters using Efficient Global Optimization (here: SPO)
- Kriging metamodels were earlier considered to be the best ones, but the Kriging variant must be chosen carefully and should be able to handle noisy landscapes!
- Additional nugget estimation resulted in better approximations of the real landscapes. This is caused by the biased distribution of sequential design points, leading to too small correlation lengths of the model predictions.
- Expected Improvement (EI) as infill criterion can enable more exploration, Forrester's Re-interpolating technique makes the use of EI possible also for noisy optimization problems

Outlook

- Comparison of EGO variants like interpolating, non-interpolating and EI-based EGO to get a better idea of the model quality for ML tasks
- Taking into account Forrester's Re-interpolation technique to handle noisy observations
- Perform repeated evaluations of similar design points. Instead of aggregating the points as was done before, the variance information can be used by the Kriging models, improving the estimated nugget size
- Optimal computing budget allocation by Chen [CDCY97] is a possibility to adapt the number of repetitions automatically.

Thanks for your attention!

Any questions?

T. Bartz-Beielstein, C.W.G. Lasarczyk, and M. Preuß.

Sequential parameter optimization.
In *The 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 773–780. IEEE, 2005.

H.C. Chen, L. Dai, C.H. Chen, and E. Yucesan.

New development of optimal computing budget allocation for discrete event simulation.
In *Simulation Conference*, pages 334–341. IEEE, 1997.

R.A. Fisher.

Design of experiments.
British Medical Journal, 1(3923):554–554, 1936.

A.I.J. Forrester, A.J. Keane, and N.W. Bressloff.

Design and analysis of "noisy" computer experiments.
AIAA journal, 44(10):2331, 2006.

N. Hansen and A. Ostermeier.

Completely derandomized self-adaptation in evolution strategies.
Evolutionary Computation, 9:159–195, 2001.

D.R. Jones, M. Schonlau, and W.J. Welch.

Efficient global optimization of expensive black-box functions.
Journal of Global optimization, 13(4):455–492, 1998.

W. Konen, P. Koch, O. Flasch, T. Bartz-Beielstein, Martina Friese, and Boris Naujoks.

Tuned data mining: A benchmark study on different tuners.
In Natalio Krasnogor, editor, *GECCO '11: Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011.

T.G. Kolda, R.M. Lewis, and V. Torczon.

Optimization by direct search: New perspectives on some classical and modern methods.
SIAM review, pages 385–482, 2003.

Wolfgang Konen.

The TDM framework: Tuned data mining in R.

CIOP Technical Report 01-11, Cologne University of Applied Sciences, Jan 2011.

D.G. Krige.

A statistical approach to some basic mine valuation problems on the witwatersrand.
Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52(6):119–139, 1951.

G. Matheron.

Principles of geostatistics.
Economic geology, 58(8):1246–1266, 1963.

M. Schonlau, W.J. Welch, and D.R. Jones.

Global versus local search in constrained optimization of computer models.
Lecture Notes-Monograph Series, pages 11–25, 1998.

J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn.

Design and analysis of computer experiments.
Statistical science, 4(4):409–423, 1989.

T. Wagner.

A subjective review of the state of the art in model-based parameter tuning.
In *Workshop on Experimental Methods for the Assessment of Computational Systems (WEMACS 2010), held in conjunction with the 11th Conference on Parallel Problem Solving from Nature (PPSN) 2010*, pages 1–13, 2010.