Selection of a covariance kernel for a Gaussian random field aimed for modeling global optimization problems

Antanas Žilinskas[1,a)], Anatoly Zhigljavsky[2,b)], Vladimir Nekrutkin[3,c)] and Vladimir Kornikov[3,d)]

[1]*Vilnius University, Institute of Data Science and Digital Technologies, Akademijos 4, LT-08663 Vilnius, Lithuania*
[2]*School of Mathematics, Cardiff University, Cardiff CF24 1AG, UK*
[3]*St. Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034 Russia.*

[a)]antanas.zilinskas@mii.vu.lt
[b)]ZhigljavskyAA@cardiff.ac.uk
[c)]vnekr@statmod.ru
[d)]vkornikov@mail.ru

Abstract. Bayesian approach is actively used to develop global optimization algorithms aimed at expensive black box functions. One of the challenges in this approach is the selection of an appropriate model for the objective function. Normally, a Gaussian random field is chosen as a theoretical model. However, the problem of estimation of parameters, using objective function values, is not thoroughly researched. In this paper, we consider the behavior of maximum likelihood estimators (MLEs) of parameters of the homogeneous isotropic Gaussian random field with squared exponential covariance function. We also compare properties of exponential covariance function models.

The model

In Bayesian approach for stochastic global optimization, the objective function f is assumed to be a realization of a Gaussian random field with a specified covariance, see [3, 9, 11]. We assume that f is modelled by a Gaussian homogeneous isotropic random field $\xi(x)$, $x \in [0,1]^d$, so that $\mathbb{E}\xi(x) = \mu$ and

$$K_\gamma(x,x') = \mathrm{Cov}(\xi(x),\xi(x')) = \mathbb{E}(\xi(x)-\mu)(\xi(x')-\mu) = \sigma^2 \prod_{i=1}^{d} \exp\{-\lambda|x_i - x_i'|^\gamma\}\,. \tag{1}$$

The parameters of the model are: $\mu \in \mathbb{R}$, $\sigma^2 \geq 0$, $\lambda > 0$ and $\gamma \in (0,2]$. The model (1) is widely used in stochastic global optimization and computer experiments, see e.g. [5, 6]. The most influential parameter is γ; this is the reason why we use the notation K_γ for the covariance kernel (1).

Note that besides (1), many other covariance kernels have been considered in the literature on machine learning and computer experiments, see [4] and [7]. The most popular families of kernels are (1) and the family of Matérn kernels, which is considered in detail in [4].

Approximating K_2 by $K_{2-\varepsilon}$ with small ε

The parameter γ determines the smoothness of realizations of the random field with covariance function (1). We distinguish three situations: (i) for $\gamma \in (0,1)$ the realizations are not differentiable, (ii) for $\gamma \in [1,2)$ the realizations are exactly once differentiable in each direction with the derivatives satisfying Holder condition with parameter $\gamma - 1$, and (iii) for $\gamma = 2$ the realizations are differentiable infinitely many times. In view of this and since a typical objective function is quite smooth, it is tempting to assume $\gamma = 2$. This assumption, however, leads to difficult mathematical

and computational problems which we discuss below. Note that the use of the model described by the kernel $K_2(x, x')$ has been already received some critics in the literature on computer experiments, see e.g. [7].

Assume $d = 1$, $\sigma^2 = 1$, $\lambda = 1$. Let us discretize the interval $[0, 1]$ and replace the random process $\xi(x)$, $x \in [0, 1]$ with an N-dimensional random vector $\xi_N = (\xi(x_1), \dots, \xi(x_N))^T$, where $x_i = (i - 0.5)/N \in [0, 1]$, $i = 1, \dots, N$. The mean of ξ_N is $\mu 1_N$, where $1_N = (1, 1, \dots, 1)^T$ is N-vector of ones; the covariance matrix of ξ_N is $W_N = (K_\gamma(x_i, x_j))_{i,j=1}^N$.

To simulate realizations of ξ_N or to implement any version of the Bayesian optimization algorithm, one needs to invert the covariance matrix W_N. Figure 1 (left and middle) demonstrate that numerical inversion of W_N could be extremely hard for $\gamma = 2$ and rather easy for $\gamma = 2 - \varepsilon$ even if $\varepsilon > 0$ is very small. For example, for $N = 100$ we have $\log_{10}(\lambda_{\min}(W_{100}) \simeq -268.58623$ if $\gamma = 2$ and $\log_{10}(\lambda_{\min}(W_{100}) \simeq -8.36979$ if $\gamma = 1.9999$. It is rather counter-intuitive but the matrices W_N have completely different properties for $\gamma = 2$ and $\gamma = 1.9999$. On the other hand, in the range $0.5 < \gamma < 1.999$ the matrices W_N seem to have very similar properties. We illustrate this on Figure 1 (right). Summarizing this discussion we claim that substituting K_2 with $K_{2-\varepsilon}$ with small ε is wrong for two reasons: (a) the corresponding covariance matrices W_N have totally different properties, and (b) smoothness of realizations of the corresponding random processes or fields is completely different.

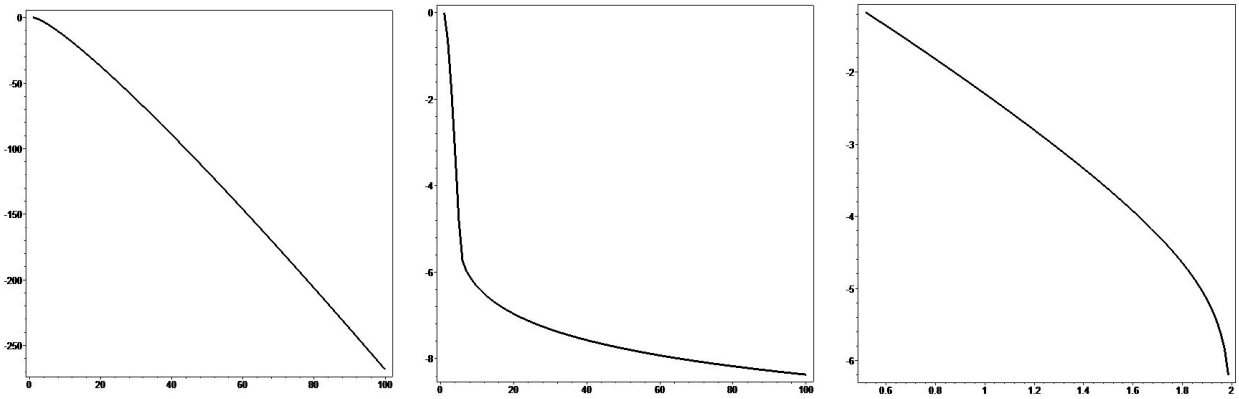

FIGURE 1. Values of $\log_{10} \lambda_{\min}$, the decimal logarithm of the minimal eigenvalue, for the matrix W_N. Left: $\gamma = 2$ and $N = 1, \dots, 100$. Middle: $\gamma = 1.9999$ and $N = 1, \dots, 100$. Right: $N = 100$ and $\gamma = 1.5, \dots, 1.99$.

Inverting the covariance matrix W_N in the case $\gamma = 2$

In this section we provide explicit expressions for the elements of the inverse of the covariance matrix for the squared exponential covariance kernel K_2; note that there are similar results obtained for the Matérn kernel with parameter $3/2$, see [1].

Assume first that $d = 1$ and the points $x_1, \dots, x_N$ are equally spaced on $[0, 1]$; for example, $x_i = (i - 0.5)/N$, $i = 1, \dots, N$. In this case, the covariance matrix W_N is $W_N = (w^{(i-j)^2})_{i,j=1}^N$ with $w = e^{-1/(\lambda_1 N^2)}$.

For two non-negative integers k and m and any real $q \neq 1$, the Gaussian binomial coefficient $\binom{k}{m}_q$ is defined by

$$\binom{k}{m}_q = \frac{(1 - q^k)(1 - q^{k-1}) \cdots (1 - q^{k-m+1})}{(1 - q)(1 - q^2) \cdots (1 - q^m)} \quad \text{for } 0 \le m \le k$$

and 0 otherwise. We shall also use the notation $C(q; i, j) = \sqrt{\prod_{k=i+1}^{j-1}(1 - q^{2k})}$. Using this notation, results of [2] imply that the matrix $W_N = (w^{(i-j)^2})_{i,j=1}^N$ has the Cholesky decomposition $W_N = LL^T$, where $L = (l_{ij})$ is lower triangular matrix with $l_{ij} = 0$ for $i < j$ and $l_{ij} = w^{(i-j)^2} C^2(w; i - j, j)/C(w; 0, j)$ for $i \ge j$. For the inverse of W_N, we have $W_N^{-1} = (L^{-1})^T L^{-1}$, where $L^{-1} = (\tilde{l}_{ij})$ with $\tilde{l}_{ij} = 0$ for $i < j$ and

$$\tilde{l}_{ij} = (-w)^{i-j} \binom{i-1}{j-1}_{w^2} \Big/ C(w; 0, i) \quad \text{for } i \ge j.$$

TABLE 1. Estimates of variance

	n=5	n=10	n=15	n=20	n=25	n=30	n=50
w	0.9608	0.9900	0.9956	0.9975	0.9984	0.9989	0.9996
p=1	0.4200	0.5477	0.5660	0.6805	0.7021	0.7973	0.9930
p=0	0.3851	0.2506	0.2117	0.1777	0.1622	0.1453	0.1127

TABLE 2. Estimates of mean and variance for the case $p = 1$

	n=5	n=10	n=15	n=20	n=25	n=30	n=50
$\hat{\mu}$	0.6000	0.5500	0.5333	0.5250	0.5200	0.5167	0.5100
$\hat{\sigma^2}$	0.2813	0.4719	0.5064	0.6315	0.6582	0.7585	0.9637

If $d > 1$ and the points x_i constitute a product grid then the covariance matrix W_N is a Kronecker product the covariance matrices in different coordinates; the same is true for the inverse of the covariance matrix, see details in [2], p.883.

MLE estimation of μ and σ^2 for $\gamma = 2$

Assume that f is a realization of a random field with mean μ and covariance kernel (1), $\gamma = 2$ and N observations of f have been made at points $x_1, \ldots, x_N$. MLE estimator of μ is $\hat{\mu} = F_N^T W_N^{-1} 1_N / 1_N^T W_N^{-1} 1_N$, where $F_N = (f(x_1), \ldots, f(x_N))^T$. The log-likelihood function of σ^2 (multiplied by 2) is

$$2L(\sigma^2) = -N \log 2\pi - N \log \sigma^2 - \log \det(W_N) - \frac{1}{\sigma^2}(F_N - \hat{\mu} 1_N)^T W_N^{-1} (F_N - \hat{\mu} 1_N)$$

and therefore the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N}\left(F_N^T W_N^{-1} F_N - \frac{(F_N^T W_N^{-1} 1_N)^2}{1_N^T W_N^{-1} 1_N}\right).$$

Some numerical results

The first step in the development of global optimization algorithms based on statistical models of objective functions is the choice of a model. A homogeneous isotropic Gaussian random field is an attractive model because of smoothness of its sampling functions. The next step is estimation of model's parameters. This step, however, is a challenging problem, as mentioned earlier. In this section we present some numerical results illustrating these challenges. The formulas defining the maximum likelihood estimates of μ and σ^2 are presented above. For the estimation of the scale parameter λ a single variable optimization problem should be solved; the algorithm described in [10] or similar one is appropriate here. The application of simply looking MLE formulas is challenging because the condition number of the correlation matrix is normally huge. Therefore the error of the computations obtained using double precision arithmetic remains unclear. This yields that the inconsistency of the testing results of the global optimization algorithms based on the statistical models can often be explained by the numerical errors in the estimation of model parameters.

In our experiments the estimates have been computed using symbolic computations of MATLAB using the formulas above. Therefore, possible inconsistencies cannot be explained by the computational errors. The sites of observations x_i are uniformly distributed in the interval [0, 1], and, to maintain comparability with [8], we consider the cases $f(x_i) = x_i^p$, where $p = 0$ and $p = 1$. In the first experiment we assume that μ and λ are known; $\mu = 0$, and the value of λ is chosen close to 1. To ensure computations in rational numbers we have to choose values of variables in the above formulas correspondingly. Let w be a rational approximation of $\exp(-1/N^2))$ implying rationality of correlations between observations at points x_i and x_j: $K_\gamma(x_i, x_j) = w^{(i-j)^2}$. The ML estimates of σ^2 based on N observations are given in Table 1. The obtained estimates behave similarly to that in [8]. Next, we assume μ unknown, and compute $\hat{\mu}$ and $\hat{\sigma}^2$ using the same data. For the case $p = 0$, the obvious result $\hat{\mu} = 0$, $\hat{\sigma^2} = 0$, is quite natural. The results for the case $p = 1$ are presented in Table 2. It looks like, that $\hat{\mu}$ converges to the arithmetic mean of observations, but $\hat{\sigma^2}$ behaves similarly to the case of $\mu = 0$ (Table 1).

TABLE 3. The second set of observations

t_i	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
y_i	0.76	0.98	0.52	-0.31	-0.93	-0.90	-0.24	0.58	0.99	0.70
t_i	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
y_i	-0.08	-0.81	-0.98	-0.46	0.38	0.95	0.85	0.15	-0.65	-1.00

Finally, we consider the case of unknown λ. The estimates are computed for two sets of observations at the uniformly distributed sites, $N = 20$. The first set of data is defined by $f(x_i) = x_i$. An increase of λ implies a monotonic increase of the likelihood, monotonic increase of $\hat{\sigma^2}$, and unchanging $\hat{\mu} = 0.525$.

The second set of data are the approximated by rational numbers values of function $f(x) = \sin(5.5\pi x))$, i.e. $f(x_i) \simeq \sin(5.5i/20)$ presented in the Table 3. For these data, the likelihood maximizer was $\hat{\lambda} = 51$ implying the following estimates of the other two parameters: $\hat{\mu} = -0.049$ and $\hat{\sigma}^2 = 0.4$. The model with the obtained estimates $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\lambda}$ seems appropriate.

Conclusions

In stochastic global optimization, the objective function is often assumed to be a realization of a Gaussian random field with squared exponential covariance function. We have shown that arising computational difficulties could be impossible to overcome unless the observations are made on a uniform grid. We also study the behavior of maximum likelihood estimators of the model, which can be very peculiar and give reasonable doubts about the adequacy of the model.

Acknowledgements

The work of A.Zilinskas was supported by the Research Council of Lithuania under Grant No. P-MIP-17-61. The work of A.Zhigljavsky was supported by a grant of Crimtant Holding Limited.

REFERENCES

[1] W.-L. Loh et al. Fixed-domain asymptotics for a subclass of matérn-type gaussian random fields. *The Annals of Statistics*, 33(5):2344–2394, 2005.

[2] W.-L. Loh and T.-K. Lam. Estimating structured correlation matrices in smooth Gaussian random field models. *The Annals of Statistics*, 28:880 – 904, 2000.

[3] J. Mockus. *Bayesian Approach to Global Optimization*. Kluwer Academic Publishers, 1988.

[4] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[5] J. Sacks, S.B. Schiller, and W.J. Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

[6] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[7] M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[8] W. Xu and M.L. Stein. Maximum likelihood estimation for smooth Gaussian random field model. *SIAM/ASA Uncertainty Quantification*, 5:138 – 175, 2017.

[9] A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization*. Springer, 2008.

[10] A. Žilinskas. Optimization of one-dimensional multimodal functions, Algorithm AS133. *Journal of Royal Statistical Society, ser.C,*, 23:367–385, 1978.

[11] A. Žilinskas and A. Zhigljavsky. Stochastic global optimization: A review on the occasion of 25 years of Informatica. *Informatica*, 27(2):229 256, 2016.