The role of the nugget term in the Gaussian process method

Andrey Pepelyshev

Abstract The maximum likelihood estimate of the correlation parameter of a Gaussian process with and without of a nugget term is studied in the case of the analysis of deterministic models.

1 Introduction

The Gaussian process method is an elegant way to analyze the results of experiments in many areas of science including machine learning (Rasmussen and Williams 2006), spatial statistics (Matheron 1973, Ripley 1981, Cressie 1993, Müller 2007), and the Bayesian analysis of computer experiments (Sacks, Welch, Mitchell, and Wynn 1989, Kennedy and O'Hagan 2001, Santner, Williams, and Notz 2003). Each area has its own specific ways of employing and interpreting the Gaussian processes. The purpose of this paper is not to give a full overview, that can be found in the above references, but to discuss some issues on the nugget term for the analysis of computer experiments.

The conception of the nugget term was first introduced in geostatistics by Matheron (1962). Roughly speaking, the variogram and covariance often show a discontinuity at the origin, termed the nugget effect. The nugget effect is considered as a random noise and may represent a measurement error or short scale variabilities. The nugget term is a well explored object in spatial statistics (Pitard 1993).

Another area of the application of Gaussian processes is the Bayesian approach developed for the analysis of computer experiments. In this approach, a so-called emulator is introduced for making probabilistic judgments on the true output of the given computer model, which is called a simulator. A Gaussian process is used for a full probabilistic specification of the emulator. Thus, the emulator is utilized to measure uncertainty of different kinds, see (Kennedy and O'Hagan 2001).

Dr A. Pepelyshev
University of Sheffield, Sheffield, S3 7RH, UK, e-mail: a.pepelyshev@sheffield.ac.uk

Formally, there is no nugget term in the Gaussian process method for the analysis of deterministic models, but the nugget term can be introduced artificially, for example, for the regularization of the inversion of a covariance matrix, see (Neal 1997) for details. Gramacy and Lee (2009) reported on the usefulness of the nugget term in their research of supercomputer experiments.

The presence of the nugget term in the Gaussian process method is natural for the analysis of stochastic and simulation models. The nugget effect may represent a measurement error or an effect of random values used inside computer models (Kleijnen 2008, Kleijnen and van Beers 2005).

The influence of the nugget term for optimal designs of experiments for a number of cases have been studied in (Zhu and Stein 2005, Stehlík, Rodríguez-Díaz, Müller, and López-Fidalgo 2008).

The present paper focuses on the Gaussian process method applied for the analysis of deterministic models. It is shown that the nugget term has a great impact on the likelihood and the estimate of correlation parameter.

2 The likelihood for a Gaussian process without the nugget term

In this section, it is shown that the likelihood of a Gaussian process has an unexpected behaviour in the analysis of non-stochastic models. More precisely, for a deterministic model of observations, the maximum likelihood estimate of the correlation parameter may tend to the infinity as the number of points increases. It means that a deterministic model is approximated by a Gaussian process with the correlation function $r(x) \approx 1$ for any x.

Indeed, let $y_i = \eta(x_i)$ be the output of the model $\eta(x)$ at the point $x_i \in [0,1]$, $i = 1,\dots,n$. Note that for a deterministic model, the replication of an observation at some point gives the same output. Without loss of generality, let $x_1 < \dots < x_n$. The likelihood for a Gaussian process with constant mean β, variance σ^2 and correlation function $r(x,\tilde{x}) = e^{-|x-\tilde{x}|/\psi}$ have the form

$$p(y|\beta,\sigma,\psi) = \frac{|R|^{-1/2}}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(y-H\beta)^T R^{-1}(y-H\beta)}$$

where $y = (y_1,\dots,y_n)^T$ is the vector of output values, $R = (r(x_i,x_j|\psi))_{i,j=1}^n$ is the correlation matrix, $H = (h(x_1),\dots,h(x_n))$, and $h(x) \equiv 1$.

The maximum likelihood (ML) estimates of β and σ have the following explicit forms

$$\hat{\beta}_{ML} = (H^T R^{-1} H)^{-1} H R^{-1} y$$

and

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(y - H\hat{\beta}_{ML})^T R^{-1}(y - H\hat{\beta}_{ML}).$$

The ML estimate of ψ can be found only numerically in the following way

$$\hat{\psi}_{ML} = \arg \max_{\psi \in (0,\infty)} p(y|\hat{\beta}_{ML}, \hat{\sigma}_{ML}, \psi).$$

After substituting and simplifying, we obtain that the estimate $\hat{\psi}_{ML}$ maximizes

$$L(\psi) = \ln \left[|R|^{-1/2} \right] - \frac{n}{2} \ln \left[(y - H\hat{\beta}_{ML})^T R^{-1} (y - H\hat{\beta}_{ML}) \right].$$

For the exponential correlation function, the inverse of matrix R admits the explicit representation $R^{-1} = V^T V$ where the matrix V is defined by

$$V = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\mu_2}{\sqrt{1-\mu_2^2}} & \frac{1}{\sqrt{1-\mu_2^2}} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\mu_3}{\sqrt{1-\mu_3^2}} & \frac{1}{\sqrt{1-\mu_3^2}} & \cdots & 0 & 0 \\ & & & & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{\mu_n}{\sqrt{1-\mu_n^2}} & \frac{1}{\sqrt{1-\mu_n^2}} \end{bmatrix},$$

$\mu_i = e^{-(x_i - x_{i-1})/\psi}$. For n equidistant points $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n$, straightforward calculation shows that

$$yR^{-1}y = \frac{y_1^2 + y_n^2}{1-\lambda^2} + \sum_{i=2}^{n-1} y_i^2 \frac{1+\lambda^2}{1-\lambda^2} - 2\sum_{i=1}^{n-1} y_i y_{i+1} \frac{\lambda}{1-\lambda^2}$$

where $\lambda = e^{-\frac{1}{(n-1)\psi}}$, and

$$|R|^{-1/2} = \frac{1}{(1-\lambda^2)^{(n-1)/2}}.$$

For the model $\eta(x) = x - 1/2$, we obtain that $\hat{\beta}_{ML} = 0$ and

$$yR^{-1}y = \frac{1}{2}\frac{1}{1-\lambda^2} + \frac{n^2-5n+6}{12(n-1)} \cdot \frac{1+\lambda^2}{1-\lambda^2} - \frac{n^2-2n-3}{6(n-1)} \cdot \frac{\lambda}{1-\lambda^2}.$$

The estimate $\hat{\psi}_{ML}$ can be found explicitly in Maple and is not presented since it is a very large expression. Applying the power series expansion, we have

$$e^{-\frac{1}{(n-1)\hat{\psi}_{ML}}} = 1 - \frac{2}{n^2} - \frac{20}{3n^2} + O\left(\frac{1}{n^3}\right) \text{ and } \hat{\psi}_{ML} = \frac{n}{2} - \frac{7}{6} - \frac{7}{18n} - \frac{17}{54n^2} + O\left(\frac{1}{n^3}\right).$$

The dependence of $\hat{\psi}_{ML}$ on n is given in Figure 1 for the model $\eta(x) = x - 1/2$ at the left part and for the model $\eta(x) = \sin(2\pi x)$ at the right part. We observe that the estimate $\hat{\psi}_{ML}$ increases almost linearly as n increases for both models.

The maximum likelihood estimate of ψ for the Gaussian correlation function $r(x, \tilde{x}) = e^{-(x-\tilde{x})^2/\psi}$ is given in Figure 2. For the model $\eta(x) = x - 1/2$ we have that $\hat{\psi}_{ML} = \infty$ for any n. Note that for the model $\eta(x) = \sin(2\pi x)$, the condition number of the correlation matrix $R(\hat{\psi}_{ML})$ is of order 10^7, 10^{14}, 10^{22}, 10^{30}, and 10^{38}

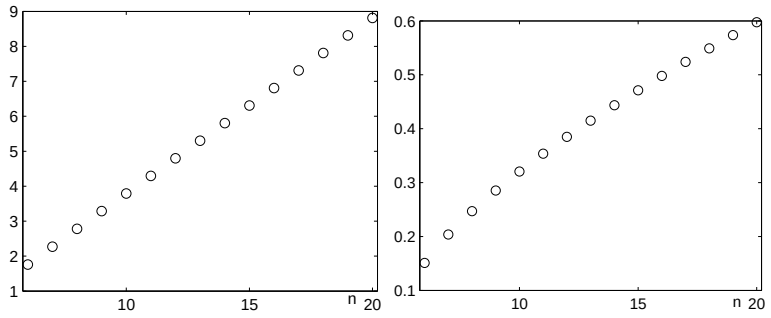

Fig. 1 The maximum likelihood estimate of ψ for the Gaussian process with the exponential correlation function and n equidistant points on the interval $[0,1]$ for the model $\eta(x) = x - 1/2$ (left part) and for the model $\eta(x) = \sin(2\pi x)$ (right part) for $n = 6, \ldots, 20$.

for $n = 8$, 11, 14, 17, and 20, respectively. These calculations were done in Maple with 45 digits precision. However, the computer representation of floating numbers typically has only 17 digits. Thus, it is impossible to find the maximum likelihood estimate for large n using the ordinary floating representation in a computer. In particular, Ababou, Bagtzoglou, and Wood (1994) have shown that the condition number grows linearly for the exponential correlation function and grows exponentially for the Gaussian correlation function.

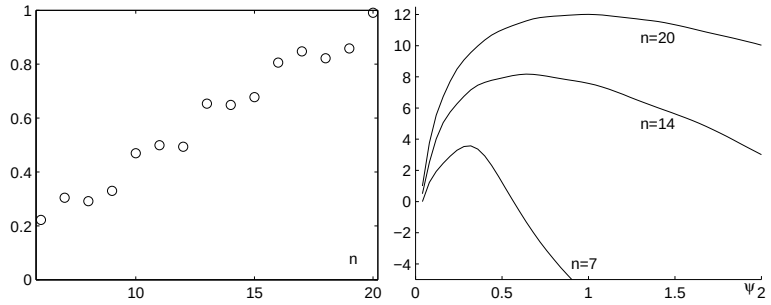

Fig. 2 At left: The maximum likelihood estimate of ψ for the Gaussian process with the Gaussian correlation function and n equidistant points on the interval $[0,1]$ for the model $\eta(x) = \sin(2\pi x)$ for $n = 6, \ldots, 20$. At right: The likelihood function of ψ for $n = 7, 14, 20$.

In more general situations for other correlation functions and other models, the dependence of the maximum likelihood estimate and the restricted maximum likelihood estimate of ψ on n remains typically the same and can be verified numerically (Pepelyshev 2009).

Thus, roughly speaking, the estimate of parameters of a Gaussian process is associated with the given data set and is not associated with the deterministic model. This estimation is not simple and is not well-defined. It is easy to observe that if one

divides an input space into several regions, one may get quite different estimates of parameters for different regions. However, if one is looking for one Gaussian process over the full space, one has difficulty in finding the single estimate.

3 The likelihood for a Gaussian process with the nugget term

3.1 MLE for a Gaussian process

In this section, the likelihood with the presence of the nugget term is investigated. For this case, the correlation matrix R in formulae from Section 2 should be replaced to the correlation matrix

$$R_\nu = \big((1-\nu)r(x_i - x_j) + \nu\delta_{i,j}\big)_{i,j}$$

where ν is the nugget term.

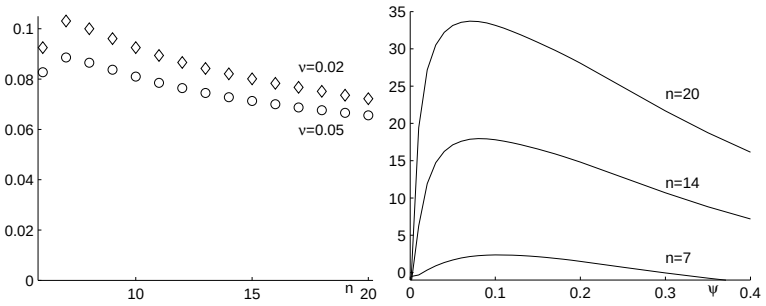

Fig. 3 At left: The maximum likelihood estimate of ψ for the Gaussian process with the Gaussian correlation function and the nugget term $\nu = 0.02, 0.05$ for measurements of the model $\eta(x) = \sin(2\pi x)$ at n equidistant points on the interval $[0,1]$, $n = 6,\ldots,20$. At right: The likelihood function of ψ for the nugget term $\nu = 0.02$ and for $n = 7, 14, 20$.

The likelihood function and the maximum likelihood estimate for fixed values of the nugget term are presented in Figure 3. One can observe that the nugget term essentially changes the maximum likelihood estimate of ψ (and also σ). The estimate $\hat{\psi}_{ML}$ does not increase to infinity as n increases, since the Gaussian process is fitted to a band around the deterministic function. It should also be noted that the condition number of the correlation matrix R_α is of order 10^2 and is increasing very slowly as n is increasing. Moreover, the estimate $\hat{\psi}_{ML}$ is smaller with the presence of the nugget term that also reduces the condition number of the correlation matrix. Ababou, Bagtzoglou, and Wood (1994) have shown that the condition number of the correlation matrix for the Gaussian process models increases to a finite limit with the presence of the nugget term.

Note one undesired effect of the nugget term. The likelihood may have the second mode for large values of the correlation parameter, see Figure 4. The second mode strongly depends on a value of the nugget term and can be considered as a false mode. For some data, the likelihood function at the second mode may have a larger value than at the first mode.

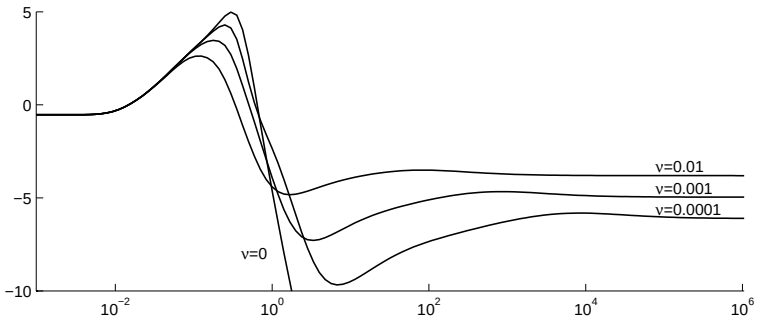

Fig. 4 The likelihood function of ψ for the Gaussian process with the Gaussian correlation function and the nugget term $v = 0, 0.01, 0.001, 0.0001$ for measurements of the model $\eta(x) = \sin(2\pi x)$ at 7 equidistant points on the interval $[0,1]$.

Note that in the presence of the nugget term, the meta-model

$$m_v(x) = H\beta + t^T(x)R_v^{-1}(y - H\beta)$$

where $t(x) = (r(x,x_1),\dots,r(x,x_n))^T$ does not possess the interpolation property. Nevertheless, the deviations $\varepsilon_i = y_i - m_v(x_i)$ are very small. One may construct a meta-model, that interpolates the dataset $\{(x_i,\varepsilon_i)\}_{i=1}^n$, by a method given in Cressie (1993, Sect. 5.9). It is not necessary for the deviations ε_i to use the Kriging approach without the nugget term. One may use the inverse distance weighted interpolation (Cressie 1993, p. 371, Lu and Wong 2008) and define the meta-model in the following form

$$m(x) = m_v(x) + \frac{\sum\limits_{i=1}^{n} \varepsilon_i ||x - x_i||_2^{-2}}{\sum\limits_{i=1}^{n} ||x - x_i||_2^{-2}}.$$

3.2 MLE for stationary processes

Let us perform a small simulation study. Assume that the results of experiments satisfy

$$y(x_i) = \beta + \sigma^2 \varepsilon^{(1)}(x_i) + \tau^2 \varepsilon^{(2)}(x_i)$$

where $x_1, \ldots, x_n$ are points of measurements, $\varepsilon^{(1)}(x)$ denotes a stationary Gaussian process with correlation function $r(x) = e^{-x^2/\psi}$ and $\varepsilon^{(2)}(x)$ is white noise. Let $\mathbf{E}\varepsilon^{(j)}(x) = 0$, $\mathbf{D}\varepsilon^{(j)}(x) = 1$, processes $\varepsilon^{(1)}(x)$ and $\varepsilon^{(2)}(x)$ be independent. The values $\beta + \sigma^2 \varepsilon^{(1)}(x_i)$ may be conceived as true values of a physical process. The values $\tau^2 \varepsilon^{(2)}(x_i)$ may be interpreted as a measurement error or a rough rounding of measured values. Let us compute the maximum likelihood estimators for 1000 realizations obtained for $n = 8$, $x_i = (i-1)/7$, $i = 1, \ldots, 8$, $\beta = 2$, $\psi = 1.5$, $\sigma = 1$, $\tau = 0$ or $\tau = 0.01$. Results of the maximum likelihood estimation for different values of the nugget term are presented in Table 1.

Table 1 The mean of maximum likelihood estimators of parameters using different values of the nugget term. Standard deviations are given in brackets.

	$\tau = 0$			$\tau = 0.01$		
ν	0	0.01	0.02	0	0.01	0.02
$\hat{\beta}_{ML}$	2.03(0.68)	2.01(0.85)	2.02(0.86)	2.02(0.92)	2.04(0.85)	2.04(0.86)
$\hat{\sigma}_{ML}$	0.83(0.40)	0.29(0.17)	0.27(0.16)	0.33(0.23)	0.30(0.17)	0.28(0.16)
$\hat{\psi}_{ML}$	1.44(0.37)	0.54(0.25)	0.47(0.20)	0.14(0.06)	0.58(0.29)	0.49(0.23)

One can observe that the maximum likelihood estimators with a nonzero nugget term does not depend on small perturbations $\{\tau^2 \varepsilon^{(2)}(x_i)\}_i$ of the data $\{\beta + \sigma^2 \varepsilon^{(1)}(x_i)\}_i$. In contrast, for $\nu = 0$, the maximum likelihood estimators of σ and ψ are significantly changed due to adding small perturbations. In all cases, the accuracy of $\hat{\beta}_{ML}$ is approximately the same. Thus, as can be seen, the nugget term yields a regularization effect on the maximum likelihood estimators.

4 Conclusions

In the analysis of deterministic models the presence of a nugget term has a significant impact on the likelihood of a Gaussian process. The maximum likelihood estimate of the correlation parameter with a nonzero nugget term is more reliable and the condition number of the correlation matrix is moderate. Even if a deterministic model does not have any internal computational errors or other perturbations, the artificial introduction of the nugget term can be recommended.

Acknowledgements Andrey Pepelyshev thanks two referees for their valuable comments and suggestions, and acknowledges the financial support provided by the MUCM project (EPSRC grant EP/D048893/1, http://mucm.group.shef.ac.uk).

References

Ababou, R., A. Bagtzoglou, and E. Wood (1994). On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology 26*(1), 99–133.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Gramacy, R. and H. Lee (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics 51*(2), 130–145.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B 63*(3), 425–450.

Kleijnen, J. P. C. (2008). *Design and analysis of simulation experiments*. International Series in Oper. Res. & Management Science, 111. New York: Springer.

Kleijnen, J. P. C. and W. C. M. van Beers (2005). Robustness of Kriging when interpolating in random simulation with heterogeneous variances: some experiments. *European J. Oper. Res. 165*(3), 826–834.

Lu, G. and D. Wong (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers and Geosciences 34*(9), 1044–1055.

Matheron, G. (1962). Traite de geostatistique appliquee. memoires bur rech. *Geol Minieres 24*.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Appl. Probability 5*, 439–468.

Müller, W. G. (2007). *Collecting spatial data* (revised ed.). Contributions to Statistics. Heidelberg: Physica-Verlag. Optimum design of experiments for random fields.

Neal, R. (1997). Monte carlo implementation of gaussian process models for bayesian classification and regression. *Technical Report 9702*.

Pepelyshev, A. (2009). Fixed-domain asymptotics of maximum likelihood estimators for observations of deterministic models. *submitted*.

Pitard, F. (1993). *Exploration of the Nugget Effect*. Doordrecht, The Netherlands: Kluwer Academic Pub.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive Comp. and Machine Learning. Cambridge: MIT Press.

Ripley, B. D. (1981). *Spatial statistics*. New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statist. Sci. 4*(4), 409–435. With comments and a rejoinder by the authors.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The design and analysis of computer experiments*. Springer Series in Statistics. New York: Springer-Verlag.

Stehlík, M., J. M. Rodríguez-Díaz, W. G. Müller, and J. López-Fidalgo (2008). Optimal allocation of bioassays in the case of parametrized covariance functions: an application to lung's retention of radioactive particles. *TEST 17*(1), 56–68.

Zhu, Z. and M. L. Stein (2005). Spatial sampling design for parameter estimation of the covariance function. *J. Statist. Plann. Inference 134*(2), 583–603.