Properties of the Posterior Distribution of a Regression Model Based on Gaussian Random Fields

A. A. Zaitsev,[*,**] **E. V. Burnaev,**[*,**,***] **and V. G. Spokoiny**[****,***]

[*] *SJC "Datadvance," Moscow, Russia*
[**] *Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia*
[***] *Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia*
[****] *Weierstrass Institute, Berlin, Germany*

Abstract—We consider the regression problem based on Gaussian processes. We assume that the prior distribution on the vector of parameters in the corresponding model of the covariance function is non-informative. Under this assumption, we prove the Bernstein–von Mises theorem that states that the posterior distribution on the parameters vector is close to the corresponding normal distribution. We show results of numerical experiments that indicate that our results apply in practically important cases.

1. INTRODUCTION

Gaussian processes are widely used to solve the regression reconstruction problem [1–3]. It is assumed that the observed sample of function values at fixed points of the design space is an implementation of a Gaussian process whose distribution is completely defined by a predefined expectation and covariance functions. It is also assumed that the covariance function between sample values depends only on the points of observation. In this case the function's value at a new point is usually predicted with posterior (with respect to the known sample of function values) expectation of the process, which is in fact a weighted sum of known values of the function, and the weights in this sum are defined by mutual covariances of function values at the new point and at sample points [1].

One usually assumes that the covariance function of a Gaussian process belongs to a certain parametric family [1] whose parameters are characterized by the prior distribution [4–6]. Accordingly, the posterior distribution of parameters (with respect to the known sample of process values) will be proportional to the product of data likelihood, which also depends on the parameters of the covariance function, and a given prior distribution parameters.

According to the well-known Bernstein–von Mises theorem (BvM), the posterior distribution is asymptotically normal whose mean is close to the maximal likelihood estimate (MLE) and the covariance matrix is close to the MLE covariance matrix. Therefore, this theorem is often considered as a Bayesian counterpart of the Fischer theorem on the MLE's asymptotic normality. The BvM theorem provides theoretical ground for various Bayesian procedures, e.g., for using Bayesian inference to find the MLE estimate and its covariance matrix, construct elliptical confidence sets based on the first two moments of the posterior distribution, and so on.

The classical version of the BvM theorem is formulated for the case when the parametric assumption regarding the data model holds and sample size tends to infinity. This is exactly the setting in which properties of the posterior distribution of the covariance function's vector of parameters are

still studied to this day (see, e.g., [7–9]). However, in practical cases it is often necessary to consider situations where sample size is limited, and the original parametric assumption on the Gaussian process covariance function may be violated (in practice it is impossible to establish the true nature of the function whose model is being constructed).

In [10, 11], the authors develop methods that help prove the BvM theorem under sufficiently general assumptions in the case of a limited sample and possible violation of the original parametric assumption regarding the model. In this work, we adapt these methods for the considered model of Gaussian processes and apply them to study the properties of the posterior distribution of the covariance function's vector of parameters. Namely, we prove the BvM theorem on the closeness of the posterior distribution of the covariance function's vector of parameters to the corresponding normal distribution in the case of a non-informative prior distribution for the vector of parameters. In particular, we show that the mean value of the posterior distribution of the vector of parameters is close to the maximal likelihood estimate (MLE), and its covariance matrix is close to the MLE covariance matrix.

The paper is organized as follows. In Section 2 we describe the regression reconstruction procedure based on Gaussian processes. In Section 3 we show our theoretical results. Section 4 presents the numerical experiments we have conducted, and the Appendix contains proofs for our theoretical results.

2. REGRESSION BASED ON GAUSSIAN PROCESSES

We solve the following problem. Consider a sample of values of an unknown function $\boldsymbol{D} = (X, \mathbf{y}) = \{\mathbf{x}_i, y(\mathbf{x}_i) = y_i\}_{i=1}^n$, $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^d$. We need to construct, given the sample $\boldsymbol{D}$ of size n, an approximation $\hat{y}(\mathbf{x})$ of the function $y(\mathbf{x})$.

We will assume that the function $y(\mathbf{x})$ is an implementation of a Gaussian process. Without loss of generality we let the mean of this Gaussian process to equal zero. In this case the joint distribution of the vector of values $\mathbf{y}$ has the form $\mathbf{y} \propto \mathcal{N}(\mathbf{0}, K)$, where K is some positive definite covariance matrix which, in general, depends on the sample $\boldsymbol{D}$.

Suppose that the covariance between arbitrary values of the Gaussian process is given by a certain covariance function $\operatorname{cov}(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$. Then the covariance matrix of sample values $\boldsymbol{D}$ has the form $K = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

For a Gaussian random process, the posterior distribution on the value of its implementation $y(\mathbf{x})$ at a new point $\mathbf{x} \in \mathbb{R}^d$ will be normal for a fixed covariance function

$$p(y(\mathbf{x})|\boldsymbol{D}) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})).$$

Expressions for the expectation $\mu(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ of the posterior distribution $p(y(\mathbf{x})|\boldsymbol{D})$ can be written explicitly as

$$\mu(\mathbf{x}) = \mathbf{k}^\top(\mathbf{x}) K^{-1} \mathbf{y},$$
$$\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^\top(\mathbf{x}) K^{-1} \mathbf{k}(\mathbf{x}).$$

Here $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^\top$ is the column vector of the covariance between the value $y(\mathbf{x})$ of the random process at point $\mathbf{x}$ and values $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n)$ of the random process in sample points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The posterior mean $\mu(\mathbf{x})$ is used as a prediction $\hat{y}(\mathbf{x})$ of the process value $y(\mathbf{x})$, and the posterior variance $\sigma^2(\mathbf{x})$ can serve as an estimate for the prediction's uncertainty.

In practice, to model a covariance function one usually uses some parametric family of covariance functions $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, where Θ is a compact set. In this case, to construct regression based on Gaussian processes it suffices to estimate the vector of parameters $\boldsymbol{\theta}$ for the covariance

function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$. Naturally, there is no reason to assume that the parametric assumption on the covariance function of a Gaussian process holds, i.e., in general $k(\mathbf{x}, \mathbf{x}') \notin \{k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'),\, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$.

The joint distribution of the vector of known values $\mathbf{y}$ will be normal. Then the logarithm of the data (quasi-) likelihood has the form

$$L(\boldsymbol{\theta}) = -\frac{1}{2}\left[n \log 2\pi + \ln |K_{\boldsymbol{\theta}}| + \mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1} \mathbf{y}\right], \tag{1}$$

where $K_{\boldsymbol{\theta}} = \{k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

As an estimate on the vector of parameters $\boldsymbol{\theta}$ one often uses the maximal (quasi-) likelihood estimate

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

Suppose that we are also given a certain prior distribution $\Pi(d\boldsymbol{\theta})$ on the vector of parameters $\boldsymbol{\theta}$. Then the posterior distribution for a given sample $\boldsymbol{D}$ will describe the conditional distribution of the random vector $\boldsymbol{\vartheta}$. This is usually written as

$$\boldsymbol{\vartheta} \,|\, \boldsymbol{D} \propto \exp\{L(\boldsymbol{\theta})\}\, \Pi(d\boldsymbol{\theta}). \tag{2}$$

The purpose of this work is to study the properties of the posterior distribution $\boldsymbol{\vartheta} \,|\, \boldsymbol{D}$. Note that the posterior distribution's maximum can be used as a characteristic value (estimate) of the vector of parameters $\boldsymbol{\theta}$.

3. PROPERTIES OF THE POSTERIOR DISTRIBUTION ON THE VECTOR OF PARAMETERS OF THE COVARIANCE FUNCTION

In what follows, we will concentrate on probabilistic properties of the posterior distribution on the vector of parameters $\boldsymbol{\theta}$ for the case of a non-informative prior distribution $\Pi(d\boldsymbol{\theta})$.

3.1. Assumptions on the Covariance Function

We introduce the following notation for the central point $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{\theta} \in \Theta} \mathrm{E} L(\boldsymbol{\theta}).$$

To describe the properties of the resulting posterior distribution, we have to impose a number of constraints on the set $X \in \mathbb{X}$, covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$, and the corresponding covariance matrices $K_{\boldsymbol{\theta}}$ and K.

Let

$$D_0^2 = -\nabla^2 \mathrm{E} L(\boldsymbol{\theta}^*), \quad V_0^2 = \mathrm{Var}\left\{\nabla L(\boldsymbol{\theta}^*)\right\}.$$

Here D_0^2 plays the role of the Fischer information matrix. Next we list the assumptions used in this work:

—*covariance function* $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ *is three times continuously differentiable with respect to* $\boldsymbol{\theta} \in \Theta$ *for* $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$;

—*minimal eigenvalues of matrices* K *and* $K_{\boldsymbol{\theta}}$ *are larger than some* $\lambda_0 > 0$*, and their maximal eigenvalues do not exceed some* $\overline{\lambda}_0 < \infty$;

—$\left\| \frac{\partial K_{\boldsymbol{\theta}}}{\partial \theta_i} \right\|_2 < \lambda_1 < \infty$ *for all* $\boldsymbol{\theta} \in \Theta,\ i = \overline{1, p}$;

—$\left\| \frac{\partial^2 K_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j} \right\|_2 < \lambda_2 < \infty$ *for all* $\boldsymbol{\theta} \in \Theta,\ i, j = \overline{1, p}$;

—$\left\| \frac{\partial^3 K_{\boldsymbol{\theta}}}{\partial\theta_i\partial\theta_j\partial\theta_k} \right\|_2 < \lambda_3 < \infty$ *for all* $\boldsymbol{\theta} \in \Theta,\ i,j,k = \overline{1,p}$;

—*the minimal eigenvalues of matrices* $\frac{1}{n}D_0^2$ *and* $\frac{1}{n}V_0^2$ *are larger than some* $d_0 > 0$ *and* $v_0 > 0$ *respectively*;

—*there exists a vector* $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}L(\boldsymbol{\theta})$;

—*there exists* $\mathtt{r}_0 > 0$ *such that for* $\forall\, r = r_0$ *and* $\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leqslant \mathtt{r}\}$ *it holds that*

$$\mathrm{E}L(\boldsymbol{\theta}) - \mathrm{E}L(\boldsymbol{\theta}^*) = \log\left(\frac{|K_{\boldsymbol{\theta}^*}|}{|K_{\boldsymbol{\theta}}|}\right) + \mathrm{tr}\left((K_{\boldsymbol{\theta}^*}^{-1} - K_{\boldsymbol{\theta}}^{-1})K\right) \neq 0.$$

Note that in this work we do not make the parametric assumption, i.e., it may happen that $k(\mathbf{x},\mathbf{x}') \notin \{k_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}'), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$.

3.2. Quadratic Exponential Covariance Function

Let us consider a sample parametric class of covariance functions, namely a quadratic exponential covariance function [1]

$$k_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\theta_i(x_i - x_i')^2\right) + \sigma^2\delta(\mathbf{x} - \mathbf{x}'), \tag{3}$$

where $\delta(\cdot)$ denotes the Kroneker function. The first term in (3) specifies the covariance between values of the Gaussian process' realizations at the points of the space, while the second term defines the variance level of the normally distributed noise in the data.

For a quadratic exponential covariance function conditions listed in Section 3.1 are assured by the choice of a sufficiently good design X and the value of the noise level $\sigma^2 \geqslant \sigma_0^2 > 0$ that plays the role of a regularization parameter in the corresponding covariance matrix $K_{\boldsymbol{\theta}}$.

In case we use this class of covariance functions we have to estimate the vector of parameters $\boldsymbol{\theta} = \{\theta_1,\ldots,\theta_p\} \in \Theta = \prod_{i=1}^{p}[\theta_{\min,i},\theta_{\max,i}]$ (here $d = p$). Note that different parameterizations of the vector of parameters $\boldsymbol{\theta}$, that have been often used [12, 13], let one improve approximation of the posterior distribution with the corresponding normal distribution.

3.3. Properties of the Posterior Distribution for the Vector of Parameters $\boldsymbol{\theta}$

We denote by $\mathtt{C}$ a universal absolute constant that in different formulas can take different values. We also fix a sufficiently large constant $\mathtt{x} = \mathtt{x}_n$, that grows as n increases.

We denote by Ω_n a *random event with dominating probability* such that

$$I\!P(\Omega_n) \geqslant 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}_n}.$$

We define the values

$$\bar{\boldsymbol{\vartheta}} \stackrel{\mathrm{def}}{=} \mathrm{E}(\boldsymbol{\vartheta} \,|\, \boldsymbol{D}), \qquad \mathfrak{S}^2 \stackrel{\mathrm{def}}{=} \mathrm{Cov}(\boldsymbol{\vartheta}) \stackrel{\mathrm{def}}{=} \mathrm{E}\left\{(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})\,(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}})^{\top} \,|\, \boldsymbol{D}\right\}$$

that play the role of posterior mean and posterior covariance of the random vector matrix $\boldsymbol{\vartheta}$.

The following generalization of the BvM theorem holds.

Theorem. *Suppose that assumptions from Section 3.1 hold. Then there exist a value* τ *and a random event* Ω_n *with dominating probability such that the following inequalities hold on* Ω_n:

$$\left\| D_0\left(\bar{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\theta}}\right)\right\|^2 \leqslant \mathtt{C}\tau(p + \mathtt{x}), \tag{4}$$

$$\left\| I\!I_p - D_0\mathfrak{S}^2 D_0\right\|_{\infty} \leqslant \mathtt{C}\tau(p + \mathtt{x}), \tag{5}$$

and the value $\tau(p + \mathtt{x})$ *is small and decreases as* n *increases.*

Besides, for an arbitrary $\boldsymbol{\lambda} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\lambda}\|^2 \leqslant (p + \mathtt{x})$ *it holds that*

$$\left|\log \mathrm{E}\left[\exp\left\{\boldsymbol{\lambda}^\top \mathfrak{S}^{-1}\left(\boldsymbol{\vartheta} - \bar{\boldsymbol{\vartheta}}\right)\right\} \mid \boldsymbol{D}\right] - \|\boldsymbol{\lambda}\|^2/2\right| \leqslant \mathtt{C}\tau(p + \mathtt{x}). \tag{6}$$

Expressions (4) and (5) show that the mean value $\bar{\boldsymbol{\vartheta}}$ and the covariance matrix $\mathfrak{S}^2$ of the posterior distribution are close to the MLE $\tilde{\boldsymbol{\theta}}$ and the D_0^{-2} matrix respectively, while Eq. (6) describes how close the posterior distribution is to the corresponding normal distribution.

4. COMPUTATIONAL EXPERIMENT

4.1. Data Generation

For simplicity we will assume that the true covariance function $k(\mathbf{x}, \mathbf{x}')$ belongs to a family of quadratic exponential covariance functions $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ (3). We will assume that noise variance is known and equals $\sigma^2 = 0.01$, while the prior distribution on the vector of parameters is uniform on a given hypercube $\Theta = \prod_{i=1}^p [\theta_{\min,i}, \theta_{\max,i}]$. This non-informative prior distribution does not distort the form of the original likelihood in the neighborhood of a point $\boldsymbol{\theta}^* \in \Theta$.

Consider a value of the vector of parameters $\boldsymbol{\theta}^*$ and a point from the set X that belongs to the hypercube $\mathbb{X} = [0, 1]^d$. Then the joint distribution on the vector of values $\mathbf{y}$ will be a multidimensional normal distribution with zero expectation and covariance matrix $K_{\boldsymbol{\theta}^*} = \{k_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

In this model, a single sample for an arbitrary $\boldsymbol{\theta} \in \Theta$ is generated as follows:

—suppose that covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ and its parameters $\boldsymbol{\theta}$ are fixed;

—generate a set of points $X = \{\mathbf{x}_i\}_{i=1}^n$ of fixed size n, e.g., with the uniform distribution on a hypercube $\mathbb{X} = [0, 1]^d$;

—generate a normally distributed vector $\mathbf{y}$ with zero expectation and covariance matrix $K_{\boldsymbol{\theta}} = \{k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ at points of X;

—the vector $\mathbf{y}$ will be a realization of the Gaussian process with fixed covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$.

4.2. The Form of the Posterior Distribution on the Data

It has been shown in [12] that there exist a covariance function and location of points in the design space such that a sample from the posterior distribution generated with the corresponding Gaussian process often has maximum at zero or infinity. Besides, [12] shows analytic examples in which the data likelihood function has a local maximum which is not global.

Figure 1 shows how the posterior distribution density depends on the value of the parameter $\boldsymbol{\theta}$. It shows two cases:

—posterior density with a single maximum located not in zero (standard case);

—posterior density has a local maximum located at zero or infinity. Note that in this case the corresponding covariance matrix violates its nondegeneracy condition.

It is clear that in the first case the posterior density is sufficiently close to normal.

Figure 2 shows an example of the posterior distribution of the vector of parameters $\boldsymbol{\theta}$ obtained in a similar way for the two-dimensional case.

4.3. Checking the Statements of the Theorem

Let us study how significant are the tails of the resulting posterior distribution. For simplicity we consider the case $d = 1$. We will be estimating the probability $\boldsymbol{\vartheta}$ to fall outside the region

$$\Theta_{\bar{\boldsymbol{\vartheta}}, 2\mathfrak{S}} = \{\boldsymbol{\theta} : |\boldsymbol{\theta} - \bar{\boldsymbol{\vartheta}}| \leqslant 2\mathfrak{S}\},$$

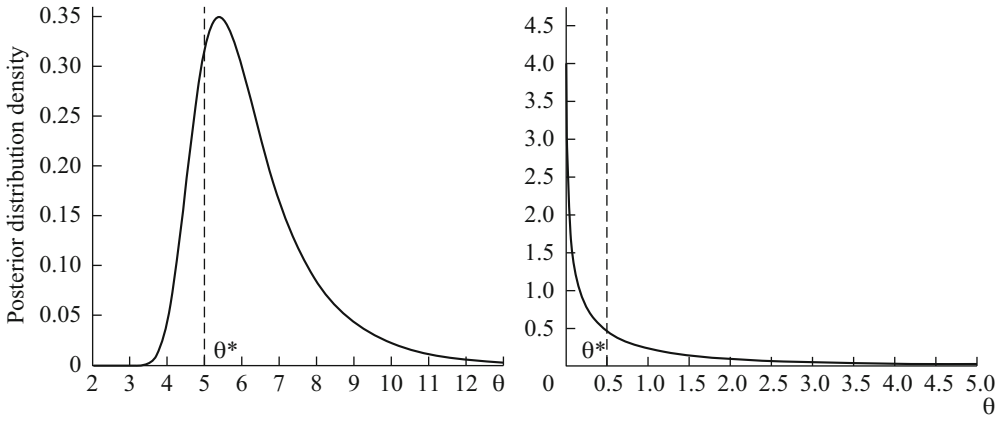

Fig. 1. Possible forms of posterior density for the distribution on the vector of parameters $\boldsymbol{\theta}$ in the one-dimensional case. On the left, the usual form of posterior density. On the right, the case when the global maximum of posterior density arises at zero. In the first case, we used a sample of size $n = 500$; in the second case, the sample size n was 50.

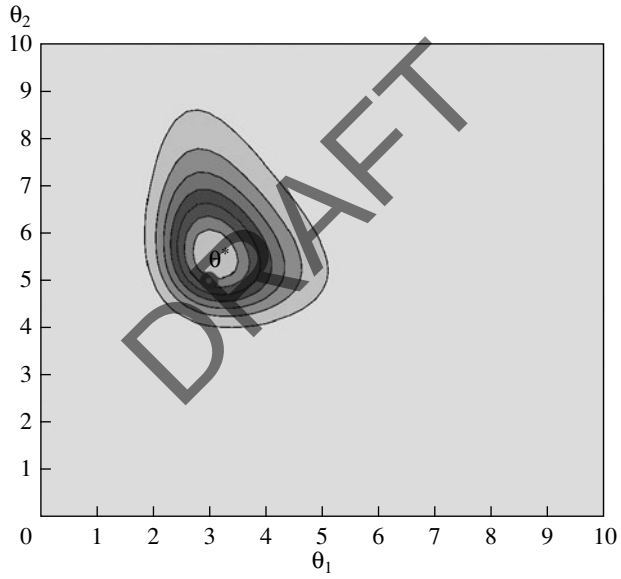

Fig. 2. Posterior distribution density for the vector of parameters $\boldsymbol{\theta}$ in the two-dimensional case. The training sample size is $n = 300$.

where $\bar{\boldsymbol{\vartheta}}$ is the expectation, and $\mathfrak{S}$ is the standard deviation of the posterior distribution.

The dependency between the probability of posterior distribution tails on the sample size is shown on Fig. 3. It is clear that as the sample size grows, the probability converges to a sufficiently small value. Confidence intervals have been estimated with bootstrapping over two hundred randomly generated samples.

For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with densities $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ respectively such that their expectations μ and variances σ^2 coincide, we define the bounded Hellinger distance with the following formula:

$$H^2(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int_{\Theta_{\mu,2\sigma}} \left(\sqrt{p(\boldsymbol{\theta})} - \sqrt{q(\boldsymbol{\theta})}\right)^2 d\boldsymbol{\theta}.$$

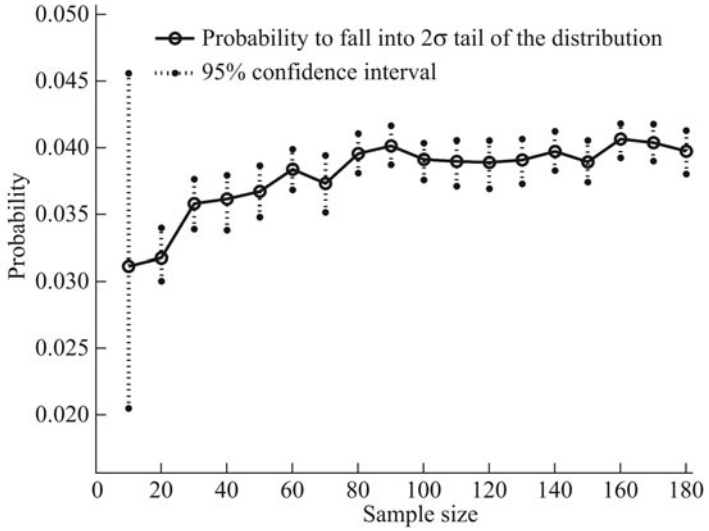

Fig. 3. Dependency of the probability to fall into the tails of the posterior distribution (solid bold line) of the sample size. Bold points denote the 95% confidence interval.

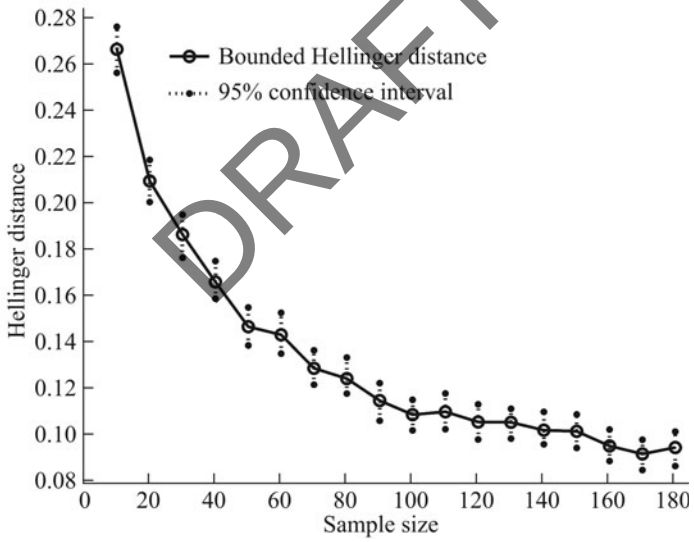

Fig. 4. Dependency of the bounded Hellinger distance (solid bold line) of the sample size. Bold points denote the 95% confidence interval.

Let us check how close posterior distribution parameters $\boldsymbol{\theta}$ are to the corresponding normal distribution in the neighborhood of the maximum point $\tilde{\boldsymbol{\theta}}$ of the posterior distribution. To do so, we compute the bounded Hellinger distance between these distributions.

Figure 4 shows that as the sample size n grows the bounded Hellinger distance between the posterior distribution and the corresponding normal distribution decreases. Confidence intervals have been estimated with bootstrapping over two hundred randomly generated samples.

Thus, indeed, as the sample size n grows both the probability of falling into posterior distribution tails and the posterior distribution converge to the corresponding normal distribution.

5. CONCLUSIONS

In this work, we have obtained a description of probabilistic properties of the posterior distribution on the vector of model parameters for the covariance function in a regression problem based on Gaussian processes. We have proven the Bernstein–von Mises theorem, namely, we have shown that the posterior distribution on the vector of parameters in case of a non-informative prior distribution is close to the corresponding normal distribution. Computational experiments show that our results are applicable in many practically important cases.

ACKNOWLEDGMENTS

This work was supported by the Laboratory for Structural Methods of Data Analysis in Predictive Modeling of the Moscow Institute of Physics and Technology (State University), grant no. 11.G34.31.0073 of the Government of Russian Federation, and the Russian Foundation for Basic Research, project no. 13-01-00521.

APPENDIX

Let us give a sketch of the proof for the theorem. We remind that $\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathrm{E}L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$. We list expressions for several values needed below:

• log-likelihood as a function of the vector of parameters $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}) = -\frac{1}{2}\left[n\log 2\pi + \ln|K_{\boldsymbol{\theta}}| + \mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1}\mathbf{y}\right];$$

• log-likelihood expectation

$$\mathrm{E}L(\boldsymbol{\theta}) = -\frac{1}{2}\left[n\log 2\pi + \ln|K_{\boldsymbol{\theta}}| + \mathrm{tr}\left(K_{\boldsymbol{\theta}}^{-1}K\right)\right];$$

• log-likelihood derivative

$$\nabla_i L(\boldsymbol{\theta}) = -\frac{1}{2}\left[\mathrm{tr}\left(K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i}\right) - \mathbf{y}^\top (K_{\boldsymbol{\theta}}^{-1})^\top \frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i}K_{\boldsymbol{\theta}}^{-1}\mathbf{y}\right];$$

• log-likelihood expectation derivative

$$\nabla_i \mathrm{E}L(\boldsymbol{\theta}) = -\frac{1}{2}\left[\mathrm{tr}\left(K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i}\right) - \mathrm{tr}\left((K_{\boldsymbol{\theta}}^{-1})^\top \frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i}K_{\boldsymbol{\theta}}^{-1}K\right)\right];$$

• Let $U_i = U_i(\boldsymbol{\theta}^*) = (K_{\boldsymbol{\theta}}^{-1})^\top \dfrac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i}K_{\boldsymbol{\theta}}^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$. The matrix $V_0^2 = \mathrm{Var}\left\{\nabla L(\boldsymbol{\theta}^*)\right\}$ has the form

$$V_0^2 = \left\{\frac{1}{2}\mathrm{tr}(U_i K U_j K)\right\}_{i,j=1}^{p};$$

• elements of matrix $D_0^2 = -\nabla^2 \mathrm{E}L(\boldsymbol{\theta}^*) = \{d_{i,j}\}_{i,j=1}^p$ have the form

$$d_{i,j} = \frac{1}{2}\,\mathrm{tr}\left[U_i K_{\boldsymbol{\theta}} U_j K + U_j K_{\boldsymbol{\theta}} U_i K - U_i K_{\boldsymbol{\theta}} U_j K_{\boldsymbol{\theta}} + (K_{\boldsymbol{\theta}}^{-1})^\top \frac{\partial^2 K_{\boldsymbol{\theta}}}{\partial\theta_i\partial\theta_j}K_{\boldsymbol{\theta}}^{-1}(K_{\boldsymbol{\theta}} - K)\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}.$$

Let us show that under the assumptions introduced in Section 3.1 the following statements hold.

Statement 1 (ED_0). *There exist constants $g > 0$, $\nu_0 \geqslant 1$ such that for all $|\lambda| \leqslant g$ it holds that*

$$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log \mathrm{E} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V_0\boldsymbol{\gamma}\|}\right\} \leqslant \nu_0^2\lambda^2/2. \tag{A.1}$$

We denote

$$Z = \frac{1}{\|V_0\boldsymbol{\gamma}\|}\sum_{i=1}^p \gamma_i U_i.$$

Then the expectation in the left-hand side of inequality (A.1) exists if matrix $K^{-1} - \lambda Z$ is positive definite or, which is the same, matrix $I - \lambda KZ$ is positive definite (here and in what follows I denotes the unit matrix of size $n \times n$). Due to our assumptions, the norm $\|KZ\|$ is bounded for arbitrary X and $\boldsymbol{\theta}^*$, therefore, there exists g such that for every $|\lambda| \leqslant g$ matrix $I - \lambda KZ \geqslant 0$.

Thus, for $|\lambda| \leqslant g$ inequality (A.1) can be rewritten as

$$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \left[-\frac{\lambda}{2}\mathrm{tr}(ZK) - \frac{1}{2}\log|I - \lambda ZK|\right] \leqslant \frac{\nu_0^2\lambda^2}{2}.$$

Since matrix $I - \lambda ZK$ is positive definite, expression under the sup sign in the left-hand side of the inequality can be decomposed into a Taylor series. As a result we get that to prove inequality (A.1) it suffices to prove that there exist ν_0^2 and g such that for every $|\lambda| \leqslant g$ it holds that

$$\left|\frac{1}{2}\mathrm{tr}\left(\sum_{i=2}^{\infty}\frac{1}{i}(\lambda ZK)^i\right)\right| \leqslant \nu_0^2\lambda^2/2. \tag{A.2}$$

Since due to the assumptions above the value $\mathrm{tr}[(ZK)]$ can be bounded from above by a certain constant c, inequality (A.2) holds for some g and ν_0^2.

Thus, we can choose parameter g so that for every $|\lambda| \leqslant g$ the following conditions hold:

(a) matrix $I - \lambda ZK$ is positive definite;

(b) the power series (A.2) whose coefficients depend on $\boldsymbol{\theta}^*$ is bounded by $\nu_0^2\lambda^2/2$.

For g chosen as above statement ED_0 holds.

Let $\mathtt{r}_0^2 \geqslant \mathtt{C}(p + \mathtt{x})$.

Statement 2 (ED_1). *For every $\mathtt{r} \leqslant \mathtt{r}_0$ there exists a constant $\omega(\mathtt{r}) \leqslant 1/2$ such that for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ it holds that*

$$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log \mathrm{E} \exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\}}{\omega(\mathtt{r})\|V_0\boldsymbol{\gamma}\|}\right\} \leqslant \nu_0^2\lambda^2/2, \qquad |\lambda| \leqslant g. \tag{A.3}$$

Here the constant g is the same as in (ED_0).

We define $Z(\boldsymbol{\theta})$ as

$$Z(\boldsymbol{\theta}) = \frac{1}{\omega(\mathtt{r})\|V_0\boldsymbol{\gamma}\|}\sum_{i=1}^p \gamma_i(U_i(\boldsymbol{\theta}^*) - U_i(\boldsymbol{\theta})),$$

where $U_i(\boldsymbol{\theta}) = \left(K_{\boldsymbol{\theta}}^{-1}\right)^\top \frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_i} K_{\boldsymbol{\theta}}^{-1}$. The proof is similar to the proof of statement (ED_0), but now we will have to additionally look for g such that for every $|\lambda| \leqslant g$ the following conditions hold:

(a) matrix $I - \lambda Z(\boldsymbol{\theta})K$ is positive definite for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$;

(b) the power series of the form (A.2) whose coefficients depend on $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ is bounded by $\nu_0^2\lambda^2/2$ for all $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$.

The constant g as above exists due to the assumptions we have introduced on the covariance function.

Statement 3 ($\mathscr{L}_0$). *For every* $\mathtt{r} \leqslant \mathtt{r}_0$ *there exists a constant* $\delta(\mathtt{r}) \leqslant 1/2$ *such that on the set* $\Theta_0(\mathtt{r})$ *it holds that*

$$\left| \frac{-2\mathrm{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} - 1 \right| \leqslant \delta(\mathtt{r}). \tag{A.4}$$

Proof of this statement uses the fact that function $\mathrm{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is twice continuously differentiable in the neighborhood of $\boldsymbol{\theta}^*$, and its gradient $\nabla \mathrm{E}L(\boldsymbol{\theta}^*)$ is zero. This means that the decomposition up to the second order $\mathrm{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ contains only the quadratic term, so the neighborhood's radius can be chosen such that inequality (A.4) holds.

Statement 4 ($\mathscr{I}$). *There exists a constant* $\mathfrak{a} > 0$ *such that*

$$\mathfrak{a}^2 D_0^2 \geqslant V_0^2. \tag{A.5}$$

Due to the properties of matrices D_0^2 and V_0^2 and assumptions we have made about the covariance function inequality (A.5) will hold.

Statement 5 ($\mathrm{E}\mathtt{r}$). *For an arbitrary* $\mathtt{r}$ *there exists* $g(\mathtt{r}) > 0$ *such that for all* $\lambda \leqslant g(\mathtt{r})$ *it holds that*

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \sup_{\boldsymbol{\gamma} \in \mathbb{R}^p} \log \mathrm{E} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})}{\|V_0 \boldsymbol{\gamma}\|} \right\} \leqslant \nu_0^2 \lambda^2 / 2. \tag{A.6}$$

Since covariance matrices and their derivatives with respect to $\boldsymbol{\theta}$ are bounded, similar to the proof of inequality (ED_0) we can show that (A.6) holds.

Statement 6 ($\mathscr{L}\mathtt{r}$). *There exists* $\mathtt{b}$ *such that for every* $\mathtt{r} \geqslant \mathtt{r}_0$

$$\inf_{\boldsymbol{\theta}:\, \|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}} \left| \mathrm{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \right| \geqslant \mathtt{b}\mathtt{r}^2. \tag{A.7}$$

Function $f(\boldsymbol{\theta}) = \frac{|\mathrm{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|}{\|V_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2}$ is continuous, and $f(\boldsymbol{\theta}) \neq 0$ for $\boldsymbol{\theta} \in \Theta_0^c(\mathtt{r}_0) = \Theta \setminus \Theta_0(\mathtt{r}_0)$ due to assumptions introduced in Section 3.1. Since $\Theta_0^c(\mathtt{r}_0)$ is a compact set, there exists $\mathtt{b} > 0$ such that $f(\boldsymbol{\theta}) \geqslant \mathtt{b}$ for $\boldsymbol{\theta} \in \Theta_0^c(\mathtt{r}_0)$, which implies (A.7).

Thus, if conditions from Section 3.1 statements (ED_0), (ED_1), ($\mathscr{L}_0$), ($\mathscr{I}$), ($\mathrm{E}\mathtt{r}$), and ($\mathscr{L}\mathtt{r}$) will also hold. The theorem is further proven with considerations shown in [11].

REFERENCES

1. Rasmussen, C.E. and Williams, C.K.I., *Gaussian Processes for Machine Learning*, Cambridge: MIT Press, 2006.
2. Chervonenkis, A.Ya., Chernova, S.S., and Zykova, T.V., Applications of Kernel Ridge Estimation to the Problem of Computing the Aerodynamical Characteristics of a Passenger Plane (in Comparison with Results Obtained with Artificial Neural Networks), *Autom. Remote Control*, 2011, vol. 72, no. 5, pp. 1061–1067.
3. Forrester, A., Sobester, A., and Keane, A., *Engineering Design via Surrogate Modelling: A Practical Guide*, Chichester: Wiley, 2008.
4. Panov, M.E., Burnaev, E.V., and Zaitsev, A.A., On Methods of Introducing Regularization in Regression Based on Gaussian processes, *Tr. konf. "Matematicheskie metody raspoznavaniya obrazov–15"* (Proc. Conf. "Mathematical Methods of Image Recognition–15"), 2011, pp. 142–145.
5. Kennedy, M.C. and Hagan, A.O., Bayesian Calibration of Computer Models, *J. R. Statist. Soc., Ser. B (Stat. Met.)*, 2001, vol. 63, no. 3, pp. 425–464.
6. Qian, P.Z.G. and Wu, C.F.G., Bayesian Hierarchical Modeling for Integrating Low-accuracy and High-accuracy Experiments, *Technometrics*, 2008, vol. 50, no. 2, pp. 192–204.

7. Kaufman, C.G., Schervish, M.J., and Nychka, D.W., Covariance Tapering for Likelihood-based Estimation in Large Spatial Data Sets, *J. Am. Statist. Ass.*, 2008, vol. 103, no. 484, pp. 1545–1555.

8. Eidsvik, J., Finley, A.O., Banerjee, S., et al., Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models, *Comput. Statist. Data Anal.*, 2011, vol. 56, no. 6, pp. 1362–1380.

9. Shaby, B. and Ruppert, D., Tapered Covariance: Bayesian Estimation and Asymptotics, *J. Comput. Graphical Statist.*, 2012, vol. 21, no. 2, pp. 433–452.

10. Spokoiny, V., Parametric Estimation. Finite Sample Theory, *Ann. Statist.*, 2012, vol. 40, no. 6, pp. 2877–2909.

11. Spokoiny, V., *Bernstein—von Mises Theorem for Growing Parameter Dimension*, preprint, 2013, arXiv: 1302.3430v2.

12. Kok, S., The Asymptotic Behaviour of the Maximum Likelihood Function of Kriging Approximations Using the Gaussian Correlation Function, *Int. Conf. on Engin. Optimization (EngOpt 2012)*, Rio de Janeiro, Brazil, 2012.

13. Nagy, B., Loeppky, J.L., and Welch, W.J., *Correlation Parameterization in Random Function Models to Improve Normal Approximation of the Likelihood or Posterior*, Technical Report, Vancouver: Univ. of British Columbia, 2007.

This paper was recommended for publication by A.V. Bernshtein, a member of the Editorial Board