Properties of the Bayesian parameter estimation of a Regression based on Gaussian Processes*

A.A. Zaytsev
(Datadvance, IITP RAS),
E.V. Burnaev, candidate phys.-math. science
(Datadvance, IITP RAS, Premolab, MIPT),
V.G. Spokoiny, doctor phys.-math. science
(Weierstrass Institute (WIAS) Berlin, Germany, Premolab, MIPT)

Abstract. We consider the regression approach based on Gaussian processes and outline our theoretical results about the properties of the posterior distribution of the corresponding covariance function's parameter vector. We perform statistical experiments confirming that the obtained theoretical propositions are valid for a wide class of covariance functions, commonly used in applied problems.

Keywords: Bayesian estimation, Gaussian Processes regression, Bernstein-von Mises theorem, posterior distribution.

1 Introduction

At the present time regression based on Gaussian Processes (GP-regression) is one of the most popular methods for recovery (approximation) of an unknown function using a sample of its values [11, 1, 7].

*This work was supported by the Laboratory for Structural Methods of Data Analysis in Predictive Modelling 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, projects no. 13-01-12447 ofi_m2 and 13-01-00521.

It is assumed that the observed sample of function values at fixed points of the design space is a realization 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 observations. In this case the function's value at a new point is usually predicted by using a posterior expectation (with respect to the known sample of function values) of the process and uncertainty of the prediction are estimated by a posterior variance. The posterior mean and the posterior variance can be calculated analytically [11] and are fully determined by the covariance function of the Gaussian process.

One usually assumes that the covariance function of a Gaussian process belongs to a certain parametric family [11] (parametric description) and therefore a GP-regression construction problem can be reduced to a problem of estimating covariance function parameters.

At the present moment theoretical results about properties of parameter estimates of a covariance function are obtained only for very special cases [8, 6, 12, 10], in particular, asymptotic framework (the sample size tends to infinity) with correctly specified parametric model is considered. At the same time a theoretical analysis of the GP-regression properties for the case of high dimensional data and finite sample size as well as the analysis of the GP-regression behaviour for the case of possible model misspecification are of vital importance. Such analysis forms grounds for justification of the often-used marginal likelihood maximisation de facto being a standard procedure in machine learning industry for estimation of GP parameters in practically important cases.

It is very natural to perform the theoretical analysis of the GP-regression method using Bayesian approach. The central result of the Bayesian statistics is the celebrated Bernstein-von Mises Theorem (BvM) about the proximity of the posterior distribution of an unknown parameter vector, defining the GP regression model, to the corresponding normal distribution.

The BvM result provides a theoretical background for different Bayesian procedures. In particularly, one can use Bayesian computations for evaluation of the MLE and its variance. Also one can build elliptic credible sets using the first two moments of the posterior, etc.

Classical asymptotic methods of statistics [5] are not suited to analyse properties of the posterior distribution for the case of growing parameter dimension and finite data sample size. Therefore new statistical approaches, based on an advanced theory of empirical processes, are necessary to perform

the analysis, see [13].

In this paper we describe our results [2, 4], which justify that for GP-regression under rather general conditions non-asymptotic version of the BvM theorem is valid for the case, when the initial parametric assumption about the GP covariance function can be not true. Obtained results provide sufficient condition on the relation between the sample size and the parameter space dimension, which guarantees the fulfilment of the BvM theorem. Results of the massive statistical modelling, being the main aim of this work, demonstrate that all the statements of the BvM theorem, proved by the authors, are valid in practice.

The paper has the following structure. In section 2 we describe procedure for function reconstruction based on Gaussian Processes. In section 3 we describe the BvM theorem. In section 4 results of performed computational experiments are given.

2 Regression based on Gaussian Processes

GP-regression is constructed as follows. Consider a sample of values of an unknown function $\mathbf{D} = (X, \mathbf{y}) = \{\mathbf{x}_i, y(\mathbf{x}_i) = y_i\}_{i=1}^n$, $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^d$. We need to construct, given the sample $\mathbf{D}$ of size n, an approximation $\widehat{y}(\mathbf{x})$ of the function $y(\mathbf{x})$.

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

Suppose that the covariance between values of the Gaussian process at arbitrary points $\mathbf{x}$ and $\mathbf{x}'$ is defined by a certain covariance function $\mathrm{cov}(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$. We denote this as $y(\mathbf{x}) \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}')\right)$. Then the covariance matrix of sample values $\mathbf{D}$ has the form $K = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

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

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

Expressions for the expectation $\mu(\mathbf{x})$ and the variance $\sigma^2(\mathbf{x})$ of the posterior

distribution $\mathrm{Law}(y(\mathbf{x})|\mathbf{D})$ 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 covariances between the value $y(\mathbf{x})$ of the random process at the point $\mathbf{x}$ and values $(y(\mathbf{x}_1),\ldots,y(\mathbf{x}_n))$ of the random process at the sample points $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. The posterior mean $\mu(\mathbf{x})$ is used as a prediction $\widehat{y}(\mathbf{x})$ of the process value $y(\mathbf{x})$, and the posterior variance $\sigma^2(\mathbf{x})$ can serve as an estimate of 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 $\boldsymbol{\theta}$ of 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 of the vector of parameters $\boldsymbol{\theta}$ one often uses the maximal (quasi-) likelihood estimate (MLE)

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}).$$

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

$$\mathrm{Law}\left(\boldsymbol{\vartheta}\,\middle|\,\mathbf{D}\right) \propto \exp\bigl\{L(\boldsymbol{\theta})\bigr\}\,\mathit{\Pi}(d\boldsymbol{\theta}). \tag{2}$$

We would like to note that the maximum of the posterior distribution can be used as a characteristic value (estimate) of the parameter vector $\boldsymbol{\theta}$ and

in the case of the non-informative prior distribution $\mathit{\Pi}(d\boldsymbol{\theta})$ it coincides with the MLE $\widetilde{\boldsymbol{\theta}}$.

The purpose of this work is to verify the theoretical results about the properties of the posterior distribution Law $\left(\boldsymbol{\vartheta} \,\middle|\, \mathbf{D}\right)$, obtained by the authors in the papers [4, 2], using statical modelling approaches.

3 Properties of the Posterior Distribution

Let us denote by $\|\cdot\|_2$ the matrix spectral norm, by $I_n \in \mathbb{R}^{n\times n}$ an identity matrix, by $\mathit{\mathbb{E}}\{\cdot\}$ a mathematical expectation and by $\mathrm{Var}\{\cdot\}$ a covariance operator with respect to the distribution $y(\mathbf{x}) \sim \mathcal{GP}\,(0, k(\mathbf{x}, \mathbf{x}'))$.

In what follows we concentrate on the case of a non-informative prior distribution $\mathit{\Pi}(d\boldsymbol{\theta})$ and possibly misspecified parametric model, i.e., in general $k(\mathbf{x}, \mathbf{x}') \notin \{k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'),\, \boldsymbol{\theta} \in \mathit{\Theta}\}$ and the underlying data distribution can lie beyond the given parametric family.

Let the following assumptions hold true:

(A1) $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ is three times continuously differentiable with respect to $\boldsymbol{\theta} \in \mathit{\Theta}$ for $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$,

(A2) $\max\{\|K\|_2, \|K_{\boldsymbol{\theta}}\|_2\} \le \overline{\lambda} < \infty$, $\max\left\{\|K^{-1}\|_2, \|K_{\boldsymbol{\theta}}^{-1}\|_2\right\} \le \lambda_0 < \infty$ for $\boldsymbol{\theta} \in \mathit{\Theta}$,

(A3) $\left\|\frac{\partial K_{\boldsymbol{\theta}}}{\partial \theta_i}\right\|_2 \le \lambda_1 < \infty$, $\left\|\frac{\partial^2 K_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j}\right\|_2 \le \lambda_2 < \infty$, $\left\|\frac{\partial^3 K_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_j \partial \theta_k}\right\|_2 \le \lambda_3 < \infty$ for $\boldsymbol{\theta} \in \mathit{\Theta}, i, j, k = \overline{1, p}$,

(A4) The central point $\boldsymbol{\theta}^* = \mathrm{argmax}_{\boldsymbol{\theta}\in\mathit{\Theta}}\, \mathit{\mathbb{E}} L(\boldsymbol{\theta})$ exists,

(A5) The minimal eigenvalues of matrices $\frac{1}{n}D_0^2$ and $\frac{1}{n}V_0^2$ are greater than $\mathsf{d}_0 > 0$ и $\mathsf{v}_0 > 0$ correspondingly, where $D_0^2 = -\nabla^2 \mathit{\mathbb{E}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and $V_0^2 = \mathrm{Var}\big\{\nabla L(\boldsymbol{\theta})\big\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.

These assumptions are analogous to the assumptions from the paper [12].

3.1 Quadratic Exponential Covariance Function

Let us consider an example of a covariance functions parametric class, namely quadratic exponential covariance function [11], which is widely used in prac-

tice [1, 7]:

$$k_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}') = \theta_1^2\left(\exp\left(-\frac{1}{2}\theta_2^2\sum_{i=1}^{d}(x_i - x_i')^2\right) + \sigma^2\delta(\mathbf{x}-\mathbf{x}')\right), \qquad (3)$$

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 design (input) space, while the second term defines the variance level of the normally distributed noise in the data.

In case of a quadratic exponential covariance function conditions, listed above, are assured by the choice of a sufficiently good (regular) design X and the value of the noise level $\sigma^2 \geq \sigma_0^2 > 0$, which plays the role of a regularization parameter for the corresponding covariance matrix $K_{\boldsymbol{\theta}}$.

3.2 The Bernstein–von Mises theorem for a Regression based on Gaussian Processes

We denote by $\mathtt{C}$ a universal absolute constant that in different formulas can take different values, by $E\{\cdot\}$ a mathematical expectation with respect to the distribution $\mathrm{Law}\big(\boldsymbol{\vartheta}\,\big|\,\mathbf{D}\big)$. We define the values $\overline{\boldsymbol{\vartheta}} \stackrel{\mathrm{def}}{=} E\big\{\boldsymbol{\vartheta}\,\big|\,\mathbf{D}\big\}, \mathfrak{S}^2 \stackrel{\mathrm{def}}{=} E\left\{\left(\boldsymbol{\vartheta}-\overline{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta}-\overline{\boldsymbol{\vartheta}}\right)^{\top}\,\big|\,\mathbf{D}\right\}$, which play the role of the posterior mean and the posterior covariance matrix of the random vector $\boldsymbol{\vartheta}$ correspondingly.

Let us also denote $\mathtt{C}_1 = \frac{4}{\mathtt{v}_0^2}\mathtt{r}_0^2\overline{\lambda}^2\lambda_0^4\left(2\lambda_1^2\lambda_0+\lambda_2\right)^2$, $\mathtt{C}_4 = \frac{4\mathtt{r}_0^2\Lambda^2}{9\mathtt{d}_0^2\mathtt{v}_0}$, $\Lambda = 4\lambda_0^6\lambda_1^3\overline{\lambda}^3 + 5.5\lambda_0^4\overline{\lambda}^2\lambda_1\lambda_2 + \overline{\lambda}\lambda_3$, where $\mathtt{r}_0 > 0$ is a constant, defined in the upper bound for an exponential moment of a derivative of the quasi log-likelihood $L(\boldsymbol{\theta})$, see [4]. The following theorem holds true [4, 2].

The Bernstein–von Mises theorem. *Suppose that assumptions (A1)–(A5) hold true and*

$$n \geq \max(\mathtt{C}_1 p, \mathtt{C}_4 p^4). \qquad (4)$$

Then there exist a value β, explicitly defined by the constants from the assumptions (A1)–(A5), and a random event Ω_n with dominating probability $\mathbb{P}\big(\Omega_n\big) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}_n}$, $\mathtt{x}_n = \log n$, such that for $\tau_n = \frac{\beta}{\sqrt{n}}$ on Ω_n

$$\left\|D_0\left(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}}\right)\right\|_2^2 \leq \mathtt{C}\tau_n(p+\mathtt{x}_n)\,, \qquad (5)$$

$$\left\|I_p - D_0\mathfrak{S}^2 D_0\right\|_{\infty} \leq \mathtt{C}\tau_n(p+\mathtt{x}_n)\,. \qquad (6)$$

Besides, for an arbitrary measurable set $A \subset I\!R^p$ *and* $\boldsymbol{\gamma} \propto \mathcal{N}(\mathbf{0}, I_p)$ *it holds that*

$$I\!P\left(\mathfrak{S}^{-1}\left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right) \in A\right) \geq e^{-\mathtt{C}\,\tau_n(p+\mathtt{x}_n)}\left\{I\!P\big(\boldsymbol{\gamma} \in A\big) - \mathtt{C}\,\tau_n(p+\mathtt{x}_n)^{1/2}\right\} - \mathtt{C}e^{-\mathtt{x}_n}, \tag{7}$$

$$I\!P\left(\mathfrak{S}^{-1}\left(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}\right) \in A\right) \leq e^{\mathtt{C}\,\tau_n(p+\mathtt{x}_n)}\left\{I\!P\big(\boldsymbol{\gamma} \in A\big) + \mathtt{C}\,\tau_n(p+\mathtt{x}_n)^{1/2}\right\} + \mathtt{C}e^{-\mathtt{x}_n}. \tag{8}$$

The main idea of the proof is to construct estimates from above for exponential moments of the quasi log-likelihood $L(\boldsymbol{\theta})$ and its derivatives and then use the results of the papers [13, 15].

Inequalities (5) and (6) show that the mean value $\overline{\boldsymbol{\vartheta}}$ and the covariance matrix $\mathfrak{S}^2$ of the posterior distribution Law $\big(\boldsymbol{\vartheta} \,\big|\, \mathbf{D}\big)$ are close to the MLE $\widetilde{\boldsymbol{\theta}}$ and the matrix D_0^{-2} correspondingly. Inequalities (7) and (8) describe how close (in total variation norm) the posterior distribution Law $\big(\boldsymbol{\vartheta} \,\big|\, \mathbf{D}\big)$ is to the corresponding normal distribution.

4 Computational Experiments

4.1 Data Generation

In experiments we use the following covariance functions:

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_1^2 \left(\exp\left(-\sum_{i=1}^{d} \theta_{i+1}^2 (x_i - x_i')^2 \right) + \sigma^2 \delta(\mathbf{x} - \mathbf{x}') \right), \tag{9}$$

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \theta_1^2 \left(\exp\left(-\theta_2^2 \sum_{i=1}^{d} (x_i - x_i')^2 \right) + \sigma^2 \delta(\mathbf{x} - \mathbf{x}') \right), \tag{10}$$

$$k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{i=1}^{d} \theta_i^2 (x_i - x_i')^2 \right) + \sigma^2 \delta(\mathbf{x} - \mathbf{x}'). \tag{11}$$

We will assume that the noise variance is known and equals $\sigma^2 = 0.001$. In the paper we assume the prior distribution on the vector of parameters to be uniform on a given hypercube $\Theta = (0, \theta_1^{\max}) \times \ldots \times (0, \theta_p^{\max})$. This non-informative prior distribution does not distort the shape of the original likelihood in the neighbourhood of a point $\boldsymbol{\theta}^*$.

Let a value of the parameter vector $\boldsymbol{\theta}^*$ is selected, the sample of points $X = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{X} = [0,1]^d$ is generated. Then the joint distribution of the vector $\mathbf{y}$ is a multidimensional normal with a zero expectation and a covariance matrix $K_{\boldsymbol{\theta}^*} = \{k_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

A single sample $\mathbf{D}$ for an arbitrary $\boldsymbol{\theta} \in \Theta$ is generated as follows:

- suppose that a 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 from 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 Form of the Data Posterior Distribution

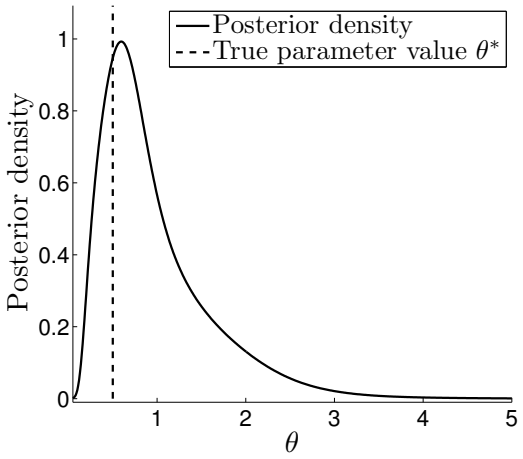
Covariance functions, widely used in practice, usually provide regular behaviour of the parameter vector posterior distribution and the corresponding quasi log-likelihood, but in some cases the quasi log-likelihood can have a maximum at zero or several local extrema [9] (for example, in case if the covariance matrix $K_{\boldsymbol{\theta}}$ is "almost" degenerate). The corresponding examples are given in figure 1.

4.3 Distribution of Parameter Estimates

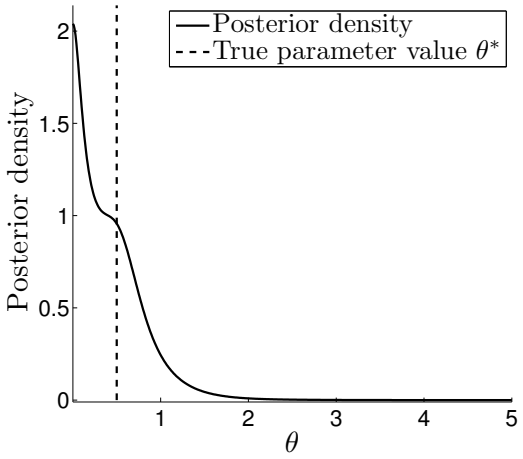
For a sample $\mathbf{D}$ we can obtain the MLE $\widetilde{\boldsymbol{\theta}}$ and the mean posterior value $\overline{\boldsymbol{\vartheta}}$. Let us investigate properties of the distributions of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\vartheta}}$ depending on the sample size.

In order to estimate a density we use a kernel density estimator with a Gaussian kernel [14]. The kernel width is selected using cross-validation procedure. Besides kernel density estimates we also show on figures 95% confidence intervals for obtained estimates.

We consider a one-dimensional case ($p = d = 1$) and a covariance function defined by (11). For $\widetilde{\boldsymbol{\theta}}$ obtained results are given in figure 2. For $\overline{\boldsymbol{\vartheta}}$ obtained

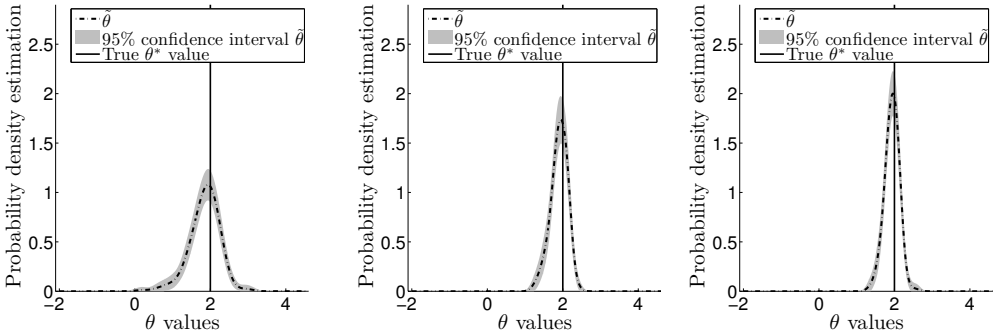


(a) Usual shape of the posterior density. The sample size is equal to $n = 50$.



(b) The global maximum of the posterior density is at zero. The sample size is equal to $n = 50$.

Figure 1: Possible shapes of the parameter vector $\boldsymbol{\vartheta}$ posterior density in one-dimensional case.



(a) The sample size n is equal to 10

(b) The sample size n is equal to 200

(c) The sample size n is equal to 1000

Figure 2: Kernel density estimate for $\widetilde{\boldsymbol{\theta}}$, parameter space dimension is equal to $p = 1$

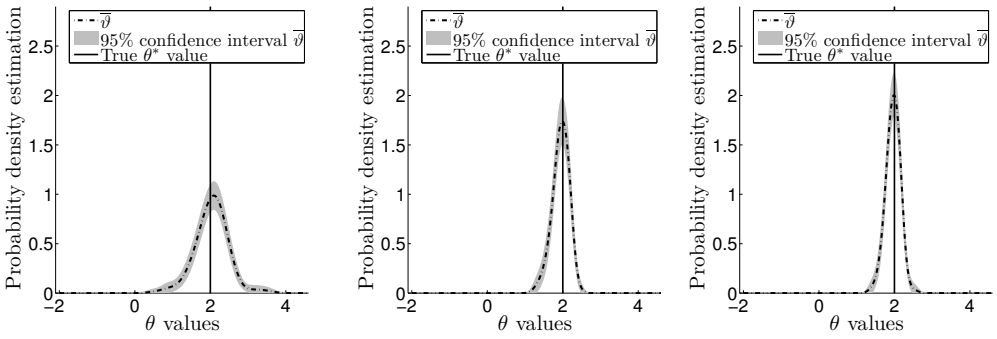
results are given in figure 3. We can see that the densities of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\vartheta}}$ concentrate around the true value when the sample size increases.

Also we consider a two-dimensional case ($p = 2$), i.e. a covariance function is defined by (10). For $\widetilde{\boldsymbol{\theta}}$ obtained results are given in figure 4. For $\overline{\boldsymbol{\vartheta}}$ obtained results are given in figure 5. In this case the densities of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\vartheta}}$ also concentrate around the true value when the sample size increases. Besides from the figures we can see that the shapes of the distributions of $\widetilde{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\vartheta}}$ are very similar.

4.4 Bound from above for deviations of the Posterior Distribution mean and covariance matrix

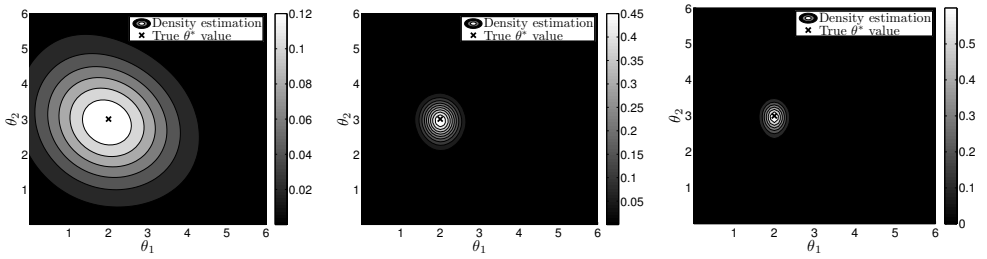
Let us consider the following experiment for one-dimensional ($p = 1$) and two-dimensional ($p = 2$) covariance functions. In theorem 3.2 it is shown that norms $\left\| D_0 \left(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}} \right) \right\|_2^2$ and $\left\| I_p - D_0 \mathfrak{S}^2 D_0 \right\|_\infty$ are bounded from above by quantities, which decreases as $\frac{1}{\sqrt{n}}$ when the sample size n increases. In figures 6 and 7 we show how these norms behaves for one-dimensional covariance function (11) and for two-dimensional covariance function (10) correspondingly. We can see that these norms both in the one-dimesional and two-dimensional cases decrease when the sample size n increases.

From figure 8 it follows that $\left\| D_0 \left(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}} \right) \right\|_2^2$ also decreases when the



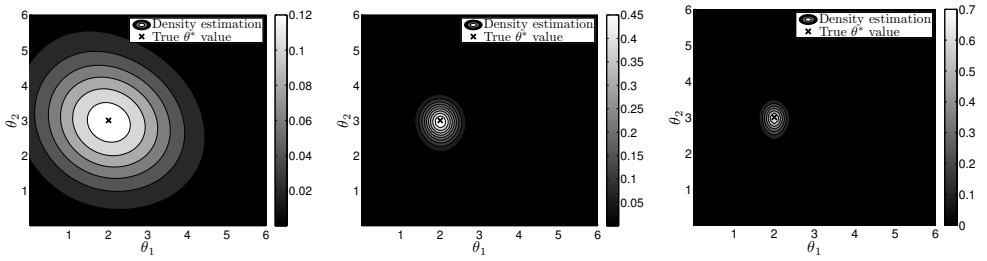
(a) The sample size n is equal to 10 (b) The sample size n is equal to 200 (c) The sample size n is equal to 1000

Figure 3: Kernel density estimate for $\overline{\boldsymbol{\vartheta}}$, parameter space dimension is equal to $p = 1$



(a) The sample size n is equal to 10 (b) The sample size n is equal to 200 (c) The sample size n is equal to 1000

Figure 4: Kernel density estimate for $\widetilde{\boldsymbol{\theta}}$, parameter space dimension is equal to $p = 2$



(a) The sample size n is equal to 10

(b) The sample size n is equal to 200

(c) The sample size n is equal to 1000

Figure 5: Kernel density estimate for $\overline{\boldsymbol{\vartheta}}$, parameter space dimension is equal to $p = 2$

sample size n increases, for example, for the case when the parameter space dimension $p = 6$.

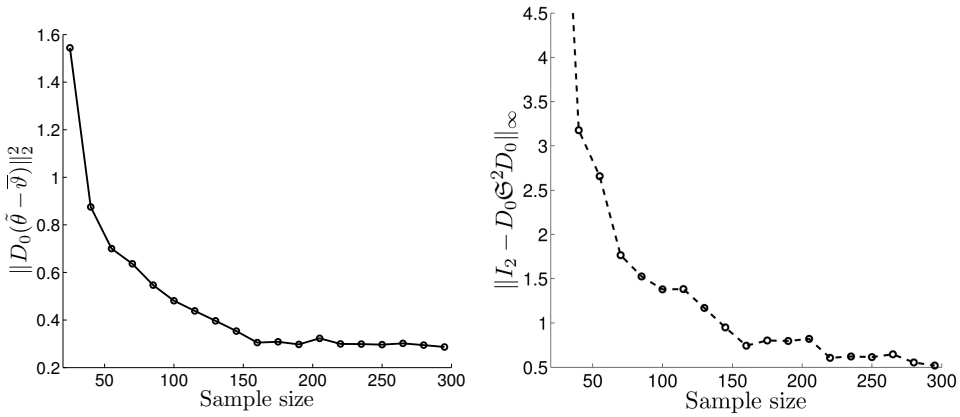
4.5 Proximity of the Posterior Distribution to the Normal Distribution

The BvM theorem states that the parameter vector $\boldsymbol{\vartheta}$ posterior distribution is close to a normal distribution: the total variation distance [3] between the posterior distribution and the corresponding normal distribution decreases when the sample size n increases. A mathematical expectation and a covariance matrix of this normal distribution are set equal to the mathematical expectation $\overline{\boldsymbol{\vartheta}}$ and the covariance matrix $\mathfrak{S}^2$ of the parameter vector $\boldsymbol{\vartheta}$ posterior distribution.

In the current section we provide experimental results for covariance function (10), which illustrates this statement of the theorem, given above.

In figure 9b we show how the Hellinger distance [3] between the considered posterior distribution and the corresponding normal distribution depends on the sample size n. In figure 9a we show how the total variation distance between the considered posterior distribution and the corresponding normal distribution depends on the sample size n.

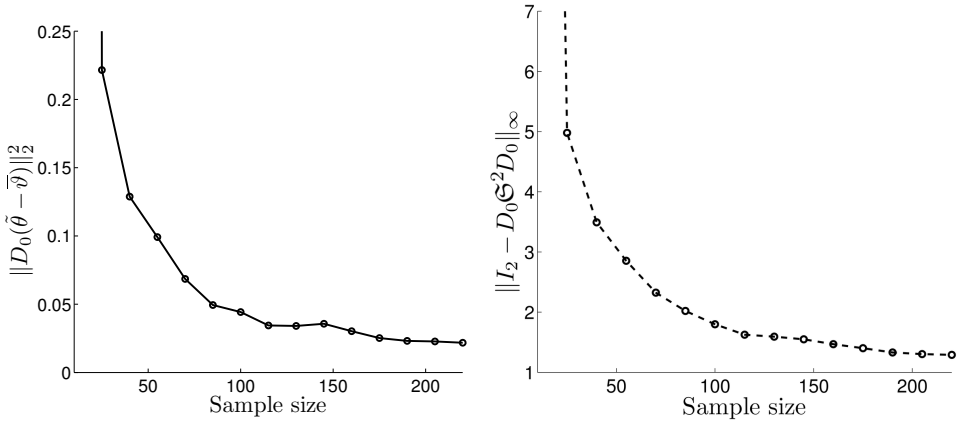
We can see that both of the distances decrease when the sample size increases. This confirms that estimates (7) and (8) are valid in the considered case.



(a) Dependence of $\left\|D_0\left(\overline{\boldsymbol{\vartheta}}-\widetilde{\boldsymbol{\theta}}\right)\right\|_2^2$ on the sample size

(b) Dependence of $\left\|I_p - D_0\mathfrak{S}^2 D_0\right\|_\infty$ on the sample size

Figure 6: Dependence of (5) and (6) on the sample size n, parameter space dimension is equal to $p = 1$



(a) Dependence of $\left\|D_0\left(\overline{\boldsymbol{\vartheta}}-\widetilde{\boldsymbol{\theta}}\right)\right\|_2^2$ on the sample size

(b) Dependence of $\left\|I_p - D_0\mathfrak{S}^2 D_0\right\|_\infty$ on the sample size

Figure 7: Dependence of (5) and (6) on the sample size n, parameter space dimension is equal to $p = 2$

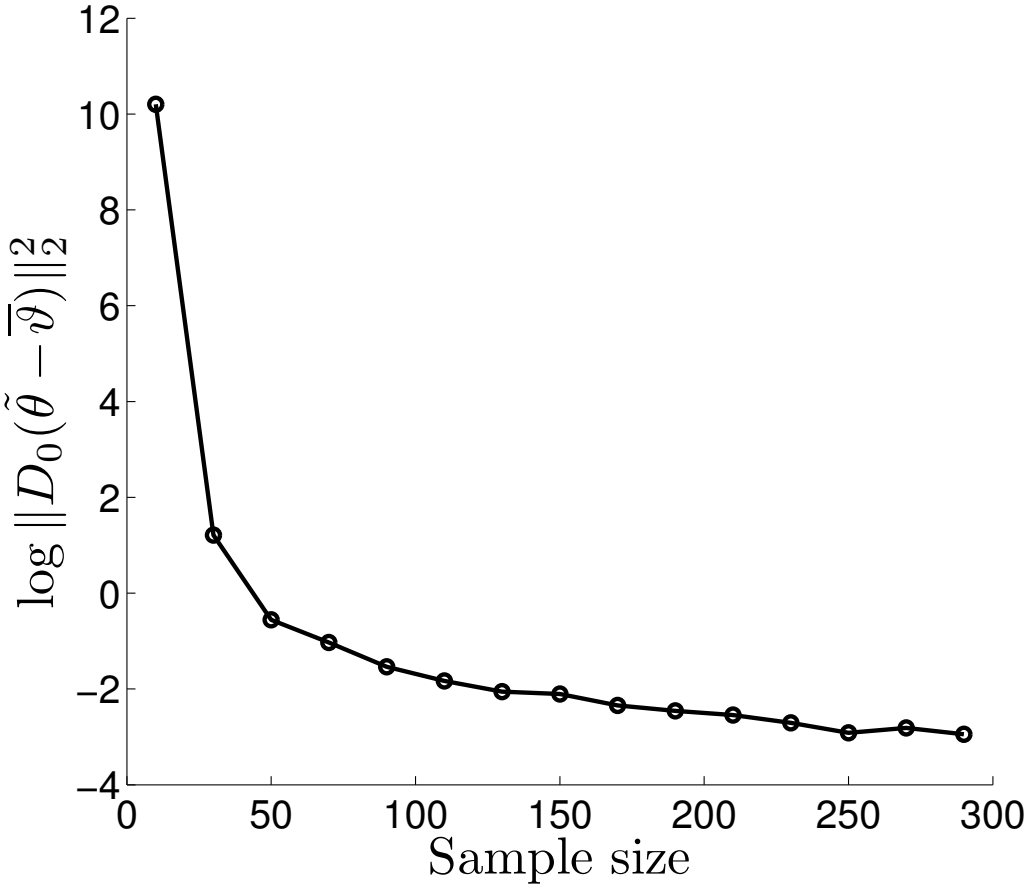
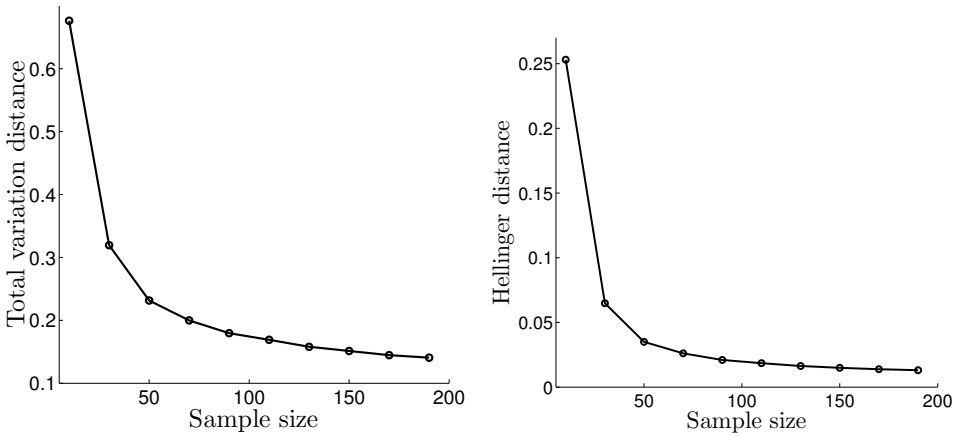


Figure 8: Dependence of $\left\| D_0 \left(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}} \right) \right\|_2^2$ on the sample size, parameter space dimension is equal to$p = 6$



(a) Total variation distance (b) Hellinger distance

Figure 9: Dependence of the distance between the parameter vector posterior distribution and the corresponding normal distribution, parameter vector dimension is equal to $p = 2$

4.6 The relation between the sample size and the dimension of the parameter space

The result, obtained in section 3.2 (see formula (4)), defines the minimal sample size n, which guarantees the fulfilment of the BvM theorem.

We consider a risk of covariance function parameter estimates. In figure 10 we show the dependence of the risk on the sample size n and the parameter space dimension p. We can see that if for each parameter space dimension p the sample size n is bigger than some critical value, then the obtained parameter estimates turn out to be precise enough.

5 Conclusions

In the paper we outline our theoretical results [2, 4], which provide grounds for the Bayesian parameter estimation of a regression based on Gaussian Processes:

- the BvM theorem for the non-asymptotic case and possibly misspecified parametric assumption about the parametric family of covariance functions,

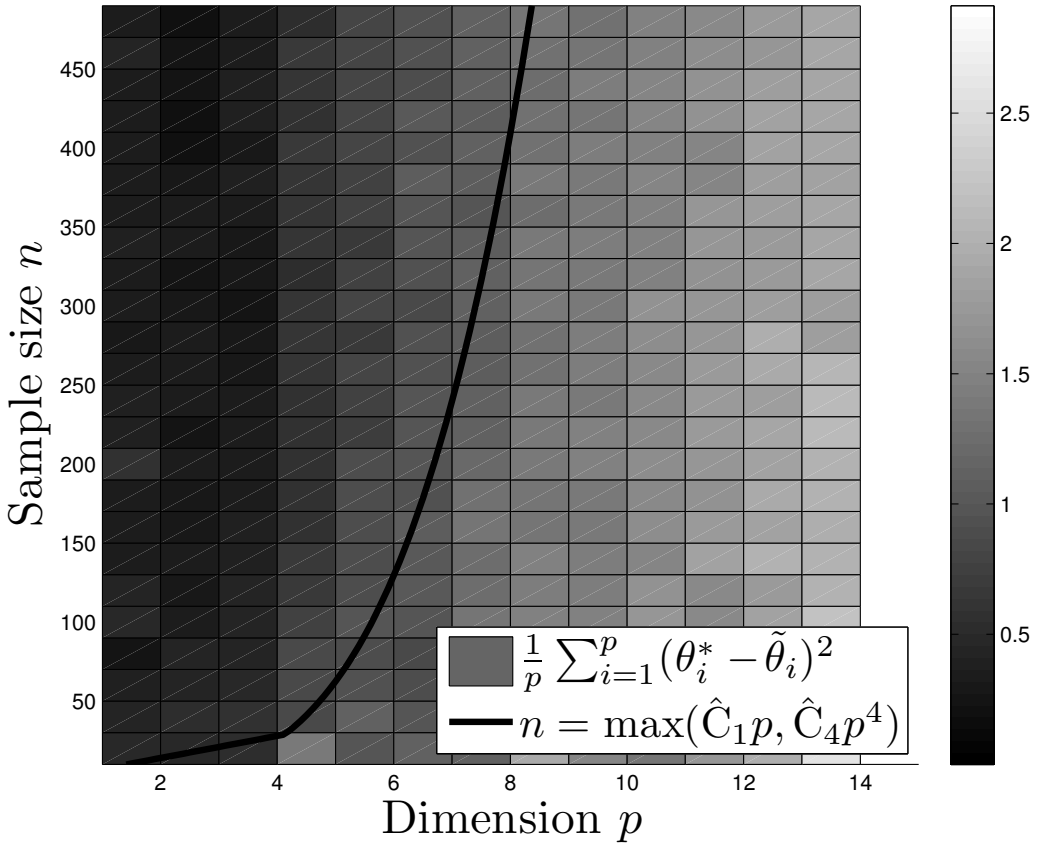


Figure 10: Dependence of the risk (of covariance function parameter estimates) on the sample size n and the parameter space dimension p

- the sufficient condition on the relation between the sample size and the dimension of the parameter space, which guarantees the fulfilment of the BvM theorem.

Results of the massive statistical modelling confirm that all statements of the BvM theorem, proved by the authors, are valid, in particular:

- in case of almost degenerate covariance matrix $K_{\boldsymbol{\theta}}$ the posterior distribution density of the parameter vector is far from the corresponding normal distribution density,
- distribution densities of the MLE $\widetilde{\boldsymbol{\theta}}$ and the posterior mean value $\overline{\boldsymbol{\vartheta}}$ concentrate around the true parameter vector $\boldsymbol{\theta}^*$,
- $\left\| D_0 \left(\overline{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\theta}} \right) \right\|_2^2$ and $\left\| I_p - D_0 \mathfrak{S}^2 D_0 \right\|_\infty$ decrease when the sample size n increases,
- the posterior distribution of the covariance function parameter vector tends (with respect to the total variation norm) to the corresponding normal distribution,
- formulated sufficient condition on the relation between the sample size n and the dimension of the parameter space p in fact influence significantly on the validity of the BvM theorem.

References

[1] A.Ya. Chervonenkis, S.S. Chernova, T.V. Zykova. 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). *Automation and Remote Control*, Volume 72, Issue 5, pp 1061-1067, May 2011.

[2] E.V. Burnaev, A.A. Zaytsev, V.G. Spokoiny, Bernstein–von Mises theorem for regression on the basis of Gaussian processes, *Uspekhi Mat. Nauk*, 68:5(413), 179–180, 2013.

[3] Shiryaev A.N. Probability (Graduate Texts in Mathematics) (v. 95) Springer, 2nd edition, 1995.

[4] A.A. Zaitsev, E.V. Burnaev, V.G. Spokoiny. Properties of the posterior distribution of a regression model based on Gaussian random fields *Automation and Remote Control*, Volume 74, Issue 10, pp. 1645-1655, October 2013.

[5] I.A. Ibragimov, R.Z. Has'minskii, Statistical estimation, asymptotic theory. Applications of Mathematics, vol. 16, Springer-Verlag, New York, 1981, vii + 403 pp.

[6] Jo Eidsvik, Andrew O Finley, Sudipto Banerjee, and Havard Rue. Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics & Data Analysis*, 56(6):1362–1380, 2011.

[7] A. Forrester, A. Sobester, and A. Keane. *Engineering design via surrogate modelling: a practical guide*. Wiley, 2008.

[8] Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

[9] S. Kok. The asymptotic behaviour of the maximum likelihood function of kriging approximations using the gaussian correlation function. In *EngOpt 2012 - International Conference on Engineering Optimization, Rio de Janeiro, Brazil, 1-5 July 2012.*, 2012.

[10] K.V. Mardia, R.J. Marshall, Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. Biometrika, 1984, Vol. 71, No. 1, P. 135–146.

[11] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[12] Benjamin Shaby and David Ruppert. Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics*, 21(2):433–452, 2012.

[13] V. Spokoiny. Parametric estimation. finite sample theory. *Annals of statistics*, 6:2877–2909, 2012.

[14] Wasserman L. All of Statistics. A Concise Course in Statistical Inference Springer, 2004.

[15] V. Spokoiny. Bernstein-von mises theorem for growing parameter dimension. *arxive.org*, 1:1, 2013.