2016 15th IEEE International Conference on Machine Learning and Applications

Conformalized Kernel Ridge Regression

Evgeny Burnaev
IITP RAS, Skoltech
Moscow, Russia
Email: e.burnaev@skoltech.ru

Ivan Nazarov
IITP RAS
Moscow, Russia
Email: ivan.nazarov@iitp.ru

***Abstract*—General predictive models do not provide a measure of confidence in predictions without Bayesian assumptions. A way to circumvent potential restrictions is to use conformal methods for constructing non-parametric confidence regions, that offer guarantees regarding validity. In this paper we provide a detailed description of a computationally efficient conformal procedure for Kernel Ridge Regression (KRR), and conduct a comparative numerical study to see how well conformal regions perform against the Bayesian confidence sets. The results suggest that conformalized KRR can yield predictive confidence regions with specified coverage rate, which is essential in constructing anomaly detection systems based on predictive models.**

***Keywords*-kernel ridge regression, gaussian process regression, conformal prediction, confidence region.**

I. Introduction

In many applied situations, like anomaly detection in telemetry of some equipment, online filtering and monitoring of potentially interesting events, or power grid load balancing, it is necessary not only to make optimal predictions with respect to some loss, but also to be able to quantify the degree of confidence in the obtained forecasts. At the same time it is necessary to take into consideration exogenous variables, that in certain ways affect the object of study.

Practical importance and difficulty of anomaly detection in general spurred a great deal of research, which resulted in a large volume of heterogeneous approaches and methods to its solution [1], [2], [3]. There are many approaches: probabilistic, which rely on approximating the generative distribution of the observed data [4], [5], metric-based anomaly detection, that measure similarity between normal and abnormal observations [6], [7], [8], predictive modelling approaches, which use the forecast error to measure abnormality [9], [10], [11], [12].

Predictive modelling is concerned with recovering an unobserved relation $x \mapsto f(x)$ from a sample $(x_i, y_i)_{i=1}^n$ of noisy observations

$$y_i = f(x_i) + \varepsilon_i, \tag{1}$$

where ε_i is iid with mean zero. One way of assigning a confidence measure to a prediction $\hat{f}$ is by using assumptions on the geometric structure of the manifold approximating the sample data which provide estimates of the likelihood of the prediction error for the test data [13], [14]. Another approach is to quantify estimate's, model's and observational uncertainty, by imposing Bayesian assumptions in (1).

Consider Kernel Ridge Regression – a model that combines ridge regression with the kernel trick. It learns a function $\hat{f}$ which, given training sample $X = (x_i)_{i=1}^n$, $X \in \mathcal{X}^{n\times 1}$, solves

$$\|y - f(X)\|^2 + \lambda \|f\|^2 \to \min_{f\in\mathcal{H}},$$

with $f(X) = (f(x_i))_{i=1}^n \in \mathbb{R}^{n\times 1}$, and $y = (y_i)_{i=1}^n \in \mathbb{R}^{n\times 1}$. Here $\mathcal{H}$ is the canonical Reproducing Kernel Hilbert Space associated with a Mercer-type kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. The Representer theorem, [15], states that the solution $\hat{f} : \mathcal{X} \mapsto \mathbb{R}$ is of the form $\hat{f}(x) = k_X'(x)\beta$ with $k_X : \mathcal{X} \mapsto \mathbb{R}^{n\times 1}$ given by the vector of canonical feature maps $k_X = (\mathcal{K}(x_i, \cdot))_{i=1}^n \in \mathcal{H}^{n\times 1}$. If e_j is the j-th unit vector in $\mathbb{R}^{n\times 1}$, and K_{XX} is the $n \times n$ Gram matrix of $\mathcal{K}$ over $(x_i)_{i=1}^n$, then $f(x_j) = k_{x_j}'\beta = e_j' K_{XX}\beta$ and $\|f\|^2 = \beta' K_{XX}\beta$. Thus the kernel ridge regression problem is equivalent to this finite-dimensional convex minimization problem:

$$\|y - K_{XX}\beta\|^2 + \lambda \beta' K_{XX}\beta \to \min_{\beta\in\mathbb{R}^{n\times 1}},$$

which yields the optimal weight vector $\hat{\beta}$ and prediction at $x^* \in \mathcal{X}$ given by $\hat{\beta} = (\lambda I_n + K_{XX})^{-1} y$ and $\hat{y}(x^*) = k_{x^*}'\hat{\beta}$, respectively.

Bayesian Kernel Ridge Regression views the model (1) as a sample path of the underlying Gaussian Process, [16], which makes predictive confidence intervals readily available. A Gaussian Process $(y_x)_{x\in\mathcal{X}}$, with mean $m : \mathcal{X} \mapsto \mathbb{R}$ and covariance kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a random process such that for any $n \geq 1$ and any $X = (x_i)_{i=1}^n \in \mathcal{X}$ the $n \times 1$ vector $y_X = (y(x_i))_{i=1}^n$ is Gaussian, $y_X \sim \mathcal{N}_n(m_X, K_{XX})$, where $m_X = (m(x_i))_{i=1}^n$. The conditional distribution of targets in a test sample $y_{X^*} = (y_{x_j^*})_{j=1}^l$ with respect to the train sample $y_X = (y_{x_i})_{i=1}^n$ is given by

$$y_{X^*}|_{y_X} \sim \mathcal{N}_l\left(m_{X^*} + K_{X^*X} Q_X (y_X - m_X), \Sigma_K(X^*)\right), \tag{2}$$

where $\Sigma_K(X^*) = K_{X^*X^*} - K_{X^*X} Q_X K_{XX^*}$, $Q_X = \left(K_{XX}\right)^{-1}$, and $K_{XX^*} = (\mathcal{K}(x_i, x_j^*)) \in \mathbb{R}^{n\times l}$. Gaussian Process Regression, or Kriging, generalizes both linear and kernel regression and assumes linearity of the mean function with respect to x and external factors h.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.65

Bayesian KRR assumes a prior on functions $f \sim GP(0, \sigma^2\mathcal{K})$ and independent Gaussian white noise $\varepsilon_x \sim \mathcal{N}(0, \sigma^2\lambda)$ in model (1), for $\sigma^2 > 0$. In this setting (2) implies that the distribution of a yet unobserved target y_{x^*} at $x^* \in \mathcal{X}$, conditional on the train data (X, y_X), is

$$y_{x^*|y_X} \sim \mathcal{N}\big(\hat{y}_{y_X}(x^*), \sigma^2\sigma_K^2(x^*)\big)\,, \tag{3}$$

with $\hat{y}_{y_X}(x^*) = k_X(x^*)' Q_X y_X$, and

$$\sigma_K^2(x^*) = \lambda + K(x^*, x^*) - k_X(x^*)' Q_X k_X(x^*)\,,$$

where $K_{XX} = (K(x_i, x_j))_{ij}$, $Q_X = \big(\lambda I_n + K_{XX}\big)^{-1}$, and $k_X = (K(x_i, \cdot))_{i=1}^n$. Thus, the $1-\alpha$ confidence interval is given by

$$\Gamma_{y_X}^{\alpha}(x^*) = \hat{y}_{y_X}(x^*) + \sigma\sqrt{\sigma_K^2(x^*)} \times [z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]\,, \tag{4}$$

where z_γ is the γ quantile of $\mathcal{N}(0,1)$. Additionally, this version naturally permits estimation of parameters of the underlying kernel $\mathcal{K}$ through maximization of the joint likelihood of the train data (X, y_X):

$$\mathcal{L} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|R_X| - \frac{1}{2\sigma^2} y' R_X^{-1} y, \tag{5}$$

where $R_X = \lambda I_n + K_{XX}$, and K_{XX} depends on the hyper-parameters of $\mathcal{K}$ (shape, precision et c.). Other approaches to estimating the covariance function's hyper-parameters are reported in [17], properties of posterior parameter distribution in Bayesian KRR are studied in [18], methods of estimating Gaussian Process Regression on large structured datasets are considered in [19], [20], and the problem of estimating in non-stationary case with regularization is considered in [21].

It is desirable to have distribution-free method that measures confidence of predictions of a machine learning algorithm. One such method is "Conformal prediction" – an approach developed in [22], which under standard independence assumptions yields a set in the space of targets, that contains yet unobserved data with a pre-specified probability. In this study, we provide empirical evidence supporting the claim that when model assumptions do hold, the conformal confidence sets, constructed over the Kernel Ridge Regression with isotropic Gaussian kernel do not perform worse than the prediction confidence intervals of a Bayesian version of the KRR. The paper is structured as follows: in section II a concise overview of what conformal prediction is and what is required to construct such kind of confidence predictor is given. Section III describes the particular steps needed to build a conformal predictor atop the kernel ridge regression. The main empirical study is reported in section IV, where we study the properties of the predictor in a batch learning setting for a KRR with specific kernel.

II. Conformal prediction

Conformal prediction is a distribution-free technique designed to yield statistically valid confidence sets for predictions made by machine learning algorithms. The key advantage of the method is that it offers coverage probability guarantees under standard iid assumptions, even in cases when assumptions of the underlying prediction algorithm fail to be satisfied. The method was introduced in [22] for online supervised and unsupervised learning.

Let $\mathcal{Z}$ denote the object-target space $\mathcal{X} \times \mathcal{Y}$. At the core of a conformal predictor is a measurable map $A : \mathcal{Z}^* \times \mathcal{Z} \mapsto \mathbb{R}$, a Non-Conformity Measure (NCM), which quantifies how much different $z_{n+1} \in \mathcal{Z}$ is relative to a sample $Z_{:n} = (z_i)_{i=1}^n \in \mathcal{Z}$. A conformal predictor over A is a procedure, which for every sample $Z_{:n}$, test object $x_{n+1} \in \mathcal{X}$, and level $\alpha \in (0,1)$, produces a confidence set $\Gamma_{Z:n}^{\alpha}(x^*)$ for the target value y_{n+1}:

$$\Gamma_{Z:n}^{\alpha}(x_{n+1}) = \big\{y \in \mathcal{Y} : p_{Z:n}(\tilde{z}_{n+1}^y) \geq \alpha\big\}\,, \tag{6}$$

where $\tilde{z}_{n+1}^y = (x_{n+1}, y)$ a synthetic test observation with target label y. The function $p : \mathcal{Z}^* \times (\mathcal{X} \times \mathcal{Y}) \mapsto [0,1]$ measures the likelihood of $\tilde{z}$ based on its non-conformity with $Z_{:n}$, and is given by

$$p_{Z:n}(\tilde{z}) = (n+1)^{-1}\big|\{i : \eta_i^{\tilde{z}} \geq \eta_{n+1}^{\tilde{z}}\}\big|\,, \tag{7}$$

where $i = 1, \ldots, n+1$, and $\eta_i^{\tilde{z}} = A(S_{-i}^{\tilde{z}}, S_i^{\tilde{z}})$ is the non-conformity of the i-th observation with respect to the augmented sample $S^{\tilde{z}} = (Z_{:n}, \tilde{z}_{n+1}^y) \in \mathcal{Z}^{n+1}$. For any i, $S_i^{\tilde{z}}$ is the i-th element of the sample, and $S_{-i}^{\tilde{z}}$ is the sample with the i-th observation omitted.

For every possible value z of an object Z_{n+1} the conformal procedure tests $H_0 : Z_{n+1} = z$, and then inverts the test to get a confidence region. The hypothesis tests are designed to have a fixed empirical type-I error rate α based on the observed sample $Z_{:n}$ and hypothesized z.

In [22] chapter 2 it has been shown, that for sequences of iid examples $(z_n)_{n\geq 1} \sim P$, the coverage probability of the prediction set (6), is at least $1-\alpha$ and successive errors are independent in online learning and prediction setting. The procedure guarantees unconditional validity: for any $\alpha > 0$

$$\mathbb{P}_{Z:n \sim P}\big(y_n \notin \Gamma_{Z:(n-1)}^{\alpha}(x_n)\big) \leq \alpha\,, \tag{8}$$

where $(x_n, y_n) = z_n$. Intuitively, the event $y_n \notin \Gamma_{Z:(n-1)}^{\alpha}(x_n)$ is equivalent to $\eta_n = A(Z_{-n}, Z_n)$ being among the largest $\lfloor n\alpha \rfloor$ values of $\eta_i = A(Z_{-i}, Z_i)$, which is equal to $n^{-1}\lfloor n\alpha \rfloor$, due to independence of $Z_{:n}$ (for a rigorous proof see [22], ch. 8).

The choice of NCM affects the size of the confidence sets and the computational burden of the conformal procedure. In the general case computing (6) requires exhaustive search through the target space $\mathcal{Y}$, which is infeasible in general regression setting. However, for specific non-conformity measures it is possible to come up with efficient procedures for computing the confidence region as demonstrated in [22] and sec. III of this paper.

III. Conformalized kernel ridge regression

In this section we describe the construction of confidence regions of the conformal procedure (6) for the case of the

non-conformity measures based on kernel ridge regression. We consider two NCMs defined in terms of regression residuals: the one used in constructing a "Ridge Regression Confidence Machine", proposed in chapter 2 of [22], and a "two-sided" NCM, proposed in [23].

A. Residuals

In each NCM it is possible to use any kind of prediction error, but we focus on the in-sample and Leave-One-Out or deleted (LOO) residuals. Consider a sample $(X,y) = (x_i, y_{x_i})_{i=1}^n$, and for any $i = 1\ldots,n$ put $X = (X_{-i}, x_i)$, and $y = (y_{-i}, y_i)$. The in-sample residuals, $\hat{r}_{\text{in}}(X,y)$, are defined for each i as

$$e_i'\hat{r}_{\text{in}}(X,y) = y_i - \hat{y}_{|(X,y)}(x_i)\,, \tag{9}$$

and LOO $\hat{r}_{\text{loo}}(X,y)$ are given by

$$e_i'\hat{r}_{\text{loo}}(X,y) = y_i - \hat{y}_{|(X_{-i},y_{-i})}(x_i)\,, \tag{10}$$

where $\hat{y}_{|(X,y)}$ and $\hat{y}_{|(X_{-i},y_{-i})}$ denote predictions of a KRR fit on the whole sample (X,y), and a sample (X_{-i},y_{-i}) with the i-th observation knocked-out, respectively. For any i the residuals are related by

$$e_i'\hat{r}_{\text{in}}(X,y) = \lambda m_i^{-1} e_i'\hat{r}_{\text{loo}}(X,y)\,, \tag{11}$$

where $\lambda m_i^{-1} = \lambda e_i' Q_X e_i$ is the KRR "leverage" score of the i-th observation, and

$$m_i = \lambda + \mathcal{K}(x_i,x_i) - k_{-i}(x_i)' Q_{-i} k_{-i}(x_i)\,, \tag{12}$$

with $k_{-i}(\cdot)$ – the $(n-1)\times 1$ vector of $(\mathcal{K}(x_j,\cdot))_{i\neq j}$, $Q_{-i} = (K_{-i} + \lambda I_{n-1})^{-1}$, and K_{-i} is the Gram matrix of the kernel $\mathcal{K}$ over subsample X_{-i}.

B. Ridge Regression Confidence Machine

In this section we describe a conformal procedure for the NMC proposed in [22] chapter 2, and focus on its in-sample version, bearing in mind that residuals (10) and (9) are interchangeable.

The Ridge Regression Confidence Machine (RRCM) constructs a non-conformity measure from the absolute value of the regression residual. The in-sample NCM, A_{in}, is given by

$$A_{\text{in}}\big((X_{-i},y_{-i}),(x_i,y_i)\big) = |e_i'\hat{r}_{\text{in}}(X,y)|\,, \tag{13}$$

and the LOO NCM, A_{loo}, is defined similarly using (10). For an NCM A the $1-\alpha$ conformal confidence interval for the n-th observation is given by

$$\Gamma^{\alpha}_{X_{-n},y_{-n}}(x_n) = \big\{z\in\mathbb{R} : p_n\big((X,\tilde{y}_n^z)\big) \geq \alpha\big\}\,, \tag{14}$$

where $\tilde{y}_i^z = (y_{-i}, z)$ – the augmented target sample y with the i-th value replaced by z. The "conformal likelihood" of the i-th observation in some sample (X,y) is given by

$$p_j\big((X,y)\big) = n^{-1}\big|\big\{j=1,\ldots,n : \eta_j \geq \eta_i\big\}\big|\,,$$

for $\eta_i = A\big((X_{-i},y_{-i}),(x_i,y_i)\big)$.

Efficient construction of the confidence set for the NCM (13) for in-sample (and deleted) residuals relies on linear dependence on the target of the n-th observation:

$$\hat{r}_i^z = e_i'\hat{r}_{\text{in}}(X,\tilde{y}_n^z) = \lambda c_i + \lambda b_i z\,, \tag{15}$$

with $c_i = e_i' C$ and $C = C_{-n}\big((X,y),x_n\big) \in \mathbb{R}^{n\times 1}$ given by

$$\begin{pmatrix} Q_{-n}y_{-n} \\ 0 \end{pmatrix} - B_{-n}(x_n) K_{-n}(x_n)' Q_{-n} y_{-n}\,,$$

where 0 is scalar and the vector $B = B_{-n}(x_n) \in \mathbb{R}^{n\times 1}$ is

$$B_{-n}(x_n) = \begin{pmatrix} -Q_{-n}K_{-n}(x_n) \\ 1 \end{pmatrix} m_n^{-1}\,. \tag{16}$$

Since absolute values of the residuals are compared, it is possible to consistently change the signs of each element of C and B to ensure that $e_i' B \geq 0$ for all i.

The conformal p-value for (x_n,y), (7), is can be re-defined in terms of regions $S_i = \{z\in\mathbb{R} : |\hat{r}_i^z| \geq |\hat{r}_n^z|\}$, for $i = 1,\ldots,n$:

$$p_{X_{-n},y_{-n}}(x_n,y) = n^{-1}\big|\{i : y\in S_i\}\big|\,. \tag{17}$$

These regions are either closed intervals, complements of open intervals, one-side closed half-rays in $\mathbb{R}$, depending on the values of C and B. In particular, with p_i and q_i denoting $-\frac{c_i+c_n}{b_i+b_n}$ and $\frac{c_i-c_n}{b_n-b_i}$, respectively (whenever each is defined), each region S_i has one of the following representations:

1) $b_i = b_n = 0$: $S_i = \mathbb{R}$ if $|c_i| \geq |c_n|$, or $S_i = \emptyset$ otherwise;
2) $b_n = b_i > 0$: S_i is either $(-\infty, p_i]$ if $c_i < c_n$, $[p_i,+\infty)$ if $c_i > c_n$, or $\mathbb{R}$ otherwise;
3) $b_n > b_i \geq 0$: S_i is either $[p_i,q_i]$ if $c_i b_n \geq c_n b_i$, or $[q_i,p_i]$ otherwise;
4) $b_i > b_n \geq 0$: S_i is $\mathbb{R}\setminus(q_i,p_i)$ when $c_i b_n \geq c_n b_i$, or $\mathbb{R}\setminus(p_i,q_i)$ otherwise.

Let P and Q be the sets of all well-defined p_i and q_i, respectively, and let $(g_j)_{j=0}^{J+1}$ enumerate distinct values of $\{\pm\infty\}\cup P\cup Q$, so that $g_j < g_{j+1}$ for all j. Then the confidence region is a closed subset of $\mathbb{R}$ constructed from sets $G_j^m = [g_j, g_{j+m}]\cap\mathbb{R}$ for $m = 0,1$:

$$\Gamma^{\alpha}_{X_{-n},y_{-n}}(x_n) = \bigcup_{m\in\{0,1\}} \bigcup_{j : N_j^m \geq n\alpha} G_j^m\,, \tag{18}$$

where $N_j^m = |\{i : G_j^m \subseteq S_i\}|$ is the coverage frequency of G_j^m, (17). In general, the resulting confidence set might contain isolated singletons G_j^0.

This set can be constructed efficiently in $\mathcal{O}(n\log n)$ time with $\mathcal{O}(n)$ memory footprint. Indeed, it is necessary to sort at most $J \leq 2n$ distinct endpoints of G_j, then locate the values p_i and q_i associated with each region S_i ($\mathcal{O}(n\log n)$). Then, since the building blocks G_j^m of Γ^{α} are either singletons ($m = 0$), or intervals made up from adjacent singletons ($m = 1$), coverage numbers N_j^m can be computed in at most $\mathcal{O}(n)$ time.

C. Kernel two-sided confidence predictor

Another possibility is to use the two-sided conformal procedure, proposed in [23]. The main result of that paper is that under relaxed Bayesian Ridge Regression assumptions if a sequence $(x_n)_{n\geq 1} \in \mathcal{X}$ is i.i.d. with an non-singular second moment matrix $\mathbb{E}x_1x_1' \succeq 0$, then for all sufficiently large n the conformal confidence regions lose little efficiency (the upper endpoints of the Bayesian and conformal prediction intervals deviate as much as $\mathcal{O}_p\big(n^{-\frac{1}{2}}\big)$).

The two-sided procedure of [23], denoted by CRR, flips the sign in (7) and uses a *conformity* measure

$$A(Z_{-i}, Z_i) = \big|\{j : \hat{r}_j \geq \hat{r}_i\}\big| \wedge \big|\{j : \hat{r}_j \leq \hat{r}_i\}\big|, \qquad (19)$$

where $(\hat{r}_i)_{i=1}^n$ are the in-sample ridge regression residuals. In that paper it was also shown that for any $\alpha \in (0,1)$ the confidence region Γ^α produced by CRR procedure for the conformity measure in (19) is equivalent to the intersection of confidence sets yielded by conformal procedures with non-conformity measures given by $\eta_i = \hat{r}_i$ and $\eta_i = -\hat{r}_i$ at significance levels $\frac{\alpha}{2}$. Individually, these NCMs define **upper** and **lower** CRR sets, and together make up a two-sided conformal procedure. Confidence regions based on this NCM, much like RRCM, can use any kind of residual: leave-one-out or in-sample.

For the upper CRR the regions $U_i = \{z \in \mathbb{R} : \hat{r}_i^z \geq \hat{r}_n^z\}$, $i = 1, \ldots, n$, are either empty, full $\mathbb{R}$ or one-side closed half-rays. Since $\hat{r}_i^z = \lambda c_i + \lambda b_i z$, U_i takes one of the following forms:

1) $b_i = b_n$: $U_i = \mathbb{R}$ if $c_i \geq c_n$, and $\emptyset$ otherwise;
2) $b_i \neq b_n$: $U_i = [q_i, +\infty)$ if $b_i > b_n$, or $U_i = (-\infty, q_i]$ otherwise;

with $q_i = \frac{c_i - c_n}{b_n - b_i}$. The forms of regions L_i for the lower CRR are computed similarly, but with the signs of c_i and b_i flipped for each $i = 1, \ldots, n$.

Both upper and lower confidence regions are built similarly to the kernel RRCM region (18) in sec. III-B. The final Kernel CRR confidence set is given by

$$\Gamma^\alpha_{X_{-n}, y_{-n}}(x_n) = \Gamma^{\alpha,\mathrm{u}}_{X_{-n}, y_{-n}}(x_n) \cap \Gamma^{\alpha,\mathrm{l}}_{X_{-n}, y_{-n}}(x_n). \qquad (20)$$

This intersection can be computed efficiently in $\mathcal{O}(n \log n)$, since the regions are built form sets anchored at a finite set Q with at most $n+2$ values. Therefore, the CRR confidence set for a fixed significance level α has $\mathcal{O}(n \log n)$ complexity.

IV. Numerical study

Validity of conformal predictors in the online learning setting has been shown in [22] chapter 2, however, no result of this kind is known in the batch learning setting. Our experiments aim to evaluate the empirical performance of the conformal prediction in this setting: with dedicated train and test datasets. In this section we conduct a set of experiments to examine the validity of the regions, produced by the conformal Kernel Ridge Regression and compare its efficiency to the Bayesian confidence intervals. We use the isotropic Gaussian kernel with the precision parameter $\theta > 0$, $\mathcal{K}(x, x') = \exp\{-\theta \|x - x'\|^2\}$, for both the Conformal Kernel ridge regression and the Gaussian Process Regression. We experiment on a compact set $\mathcal{X} \subset \mathbb{R}^{d \times 1}$, since the validity of conformal region is by design unaffected by the NCM A, which can be an arbitrary computable function and is oblivious to the structure of the domain. The dimensionality of the input data, however, may impact the width of the constructed confidence region.

The off-line validity and efficiency of Bayesian and Conformal confidence regions is studied in two settings: the fully Gaussian, and the non-Gaussian cases. In the first case the Bayesian assumptions hold and experiments are run on a sample path of $GP(0, \mathcal{K} + \delta_{x,x'}\gamma)$. In the second the assumptions of Gaussian Process Regression are partially valid: a deliberately non-Gaussian f in (1) is contaminated by moderate Gaussian white noise with variance γ.

The following hyper-parameters are controlled: the true noise-to-signal ratio $\gamma \in \{10^{-6}, 10^{-1}\}$, the covariance kernel precision $\theta \in \{10, 10^2, 10^3\}$, the train sample size n (from 25 up to 1600), the NCM – (13) or (19), the kind of residual (9), or (10), the regularization parameter $\lambda \in \{10^{-1}, 10^{-6}\}$, and either fixed θ or θ that minimizes (5).

For a given test function $f : \mathcal{X} \mapsto \mathbb{R}$ and a set of hyper-parameters each experiment consists of the following steps:

1) The test inputs, X^*, are given by a regular grid in $\mathcal{X}$ with constant spacing;
2) Train inputs, X, are sampled from a uniform distribution over $\mathcal{X}$;
3) For all $x \in X_{\mathrm{pool}} = X \cup X^*$, target values $y_x = f(x)$ are generated;
4) For $l = 1, \ldots, L$ independently:
   a) draw a random subsample of size n from train dataset;
   b) fit a Gaussian Process Regression with zero mean and Gaussian kernel $\mathcal{K}$ with the specified precision θ and λ;
   c) for each $x^* \in X^*$ construct the Bayesian (4) and conformal (6) confidence regions using the NCM A and residuals $\hat{r}$ with MLE estimated σ^2;
   d) estimate the coverage rate and the width of the convex hull of the region over the test sample X^*: $p_l(R) = |X^*|^{-1} \sum_{x \in X^*} 1_{y_x \in R_x}$, and $w_l(R) = \inf\{b - a : R \subseteq [a, b]\}$, where R is a confidence region.

A. Results: 1-d

We begin with the examination of the fully-Gaussian setup with $\mathcal{X} = [0, 1]$. To illustrate the constructed confidence regions, we generated a sample path of the 1-d Gaussian process with isotropic Gaussian kernel on a regular grid of 501 knots, and use a subset of 51 knots in $[0.05, 0.95]$ for constructing the Bayesian (GPR) and conformal (RRCM) confidence regions. The confidence regions are depicted

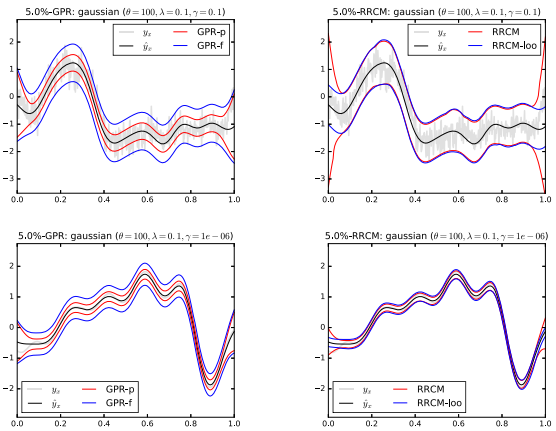


Figure 1: A sample path of a 1-d Gaussian Process with $\gamma = 10^{-1}$ (*top*), and $\gamma = 10^{-6}$ (*bottom*) constructed confidence intervals: the forecast "GPR-f" and prediction "GPR-p" (*left*), and the "RRCM" confidence bands (*right*).

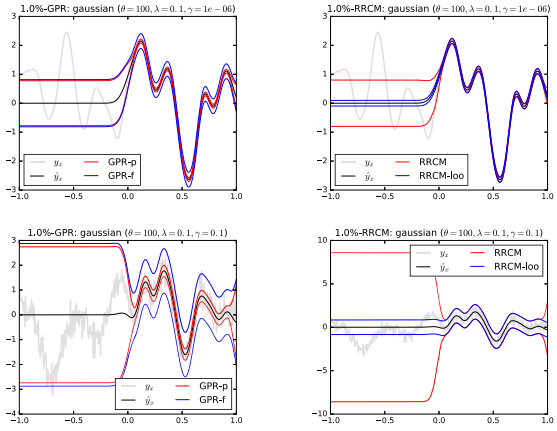


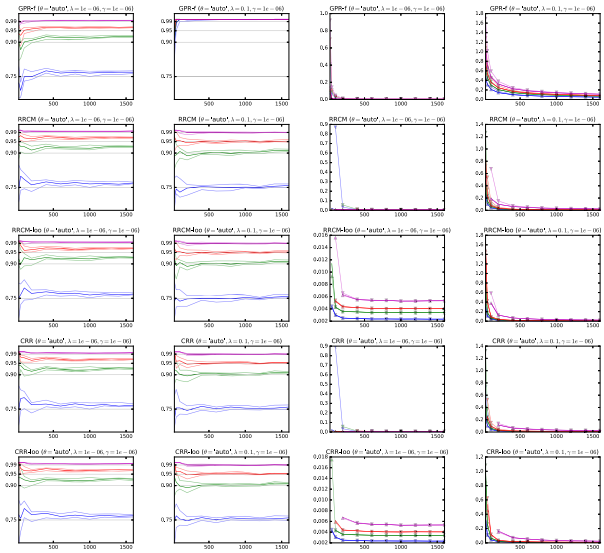
Figure 2: Limiting out-of-sample behaviour of GPR (*left*) and RRCM (*right*) confidence regions for a sample path of a Gaussian process with negligible (*top*, $\gamma = 10^{-6}$) and high (*bottom*, $\gamma = 10^{-1}$) noise-to-signal level.

in fig. 1: conformal regions closely track the Bayesian confidence bands ("GPR-f", (4)), but the latter are too wide in the low noise case. Near the endpoints the confidence regions dramatically increase in width, reflecting increased uncertainty. In general, the conformal regions necessarily cover the KRR prediction $\hat{y}^*_{|(X,y)}(x^*)$ but are not necessarily symmetric around it, where as GPR regions are (4).

In can be argued, that confidence regions for any observation x_n sufficiently far away from the bulk of the training dataset have constant size, determined only by the train sample (fig. 2). Indeed, as $\|x_n\|^2 \to \infty$ (n-fixed) the vector B_{-n} in (16) approaches the n-th unit vector e_n, since for the Gaussian kernel the vector $\|K_{-n}(x_n)\|^2 \to 0$. Since the kernel is bounded, the value m_n (12) is a bounded function of x_n, which, in turn, implies that eventually all RRCM (similarly, CRR) regions S_i assume the form of closed intervals $[-|q_i|, |q_i|]$, where $q_i = m_n(e_i' Q_{-n} y_{-n}) + o(\|x_n\|^2)$, $i \neq n$. Therefore, the conformal procedure essentially reverts to a constant-size confidence region, determined by the $n^{-1}\lfloor n(1-\alpha)\rfloor$-th order statistic of $(|q_i|)_{i=1}^n$. Analogous effects can be observed for the Gaussian Process confidence interval (4).

By construction, conformal confidence region (6) allows for some uncertainty. Indeed, the residuals (15) and hence the vector of non-conformity scores $(\eta_i^z)_{i=1}^n$ are continuous functions of targets $y = (y_i)_{i=1}^n$. Thus for small perturbations of y the relative ordering of η_i^z is kept, and the interval remains unchanged. Therefore, the set (6) tends to capture more points y for the same fixed significance level.

Experimental results in the perfectly noiseless case show that both kinds of confidence region are conservative (see fig. 3). In this picture we consider the ML estimate of θ, but the results for other fixed choices are qualitatively similar. By construction the procedure (6) adapts to the noise level in observations through the distribution of the non-



(a) (b) (c) (d)

Figure 3: Coverage rate dynamics (a, b) and asymptotic width (c, d) of the confidence regions in the fully Gaussian low-noise case $\gamma = 10^{-6}$ for $\theta = \hat{\theta}_{\mathrm{ML}}$ and $\lambda = 10^{-6}$ (a, c), and $\lambda = 10^{-1}$ (b, d). Rows from *top* to *bottom*: "GPR-f", "RRCM", "RRCM-loo", "CRR", "CRR-loo". In columns c and d upward triangles indicate the 5% sample quantile across the whole test sample, downward triangles indicate the maximal width, the median width is drawn with a slightly thicker line.

conformity scores, rather than the regularization parameter λ, which makes Bayesian confidence intervals usually wider than conformal regions (fig. 3b). Nevertheless for higher λ the coverage rate of conformal regions gets closer to the specified confidence level.

In the non-Gaussian noiseless experiment the coverage rate of conformal confidence intervals maintains the specified confidence levels and all conformal procedures demonstrate very similar asymptotic validity. For the "Heaviside"

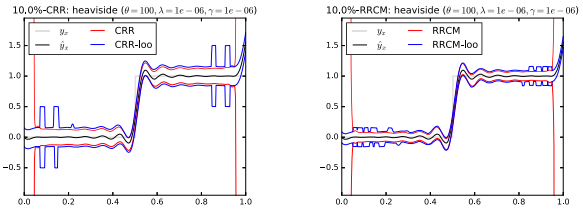


Figure 4: Typical conformal confidence bands for the "Heaviside" step function (train sample size $n = 50$): *left* – CRR, and *right* – RRCM.

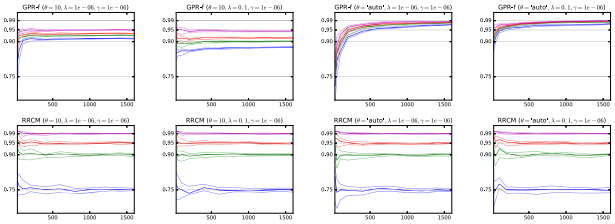


Figure 5: Coverage dynamics for the "Heaviside" ($\lambda = 10^{-6}$).

step function typical confidence bands are shown in fig. 4, and the asymptotic coverage rate of various confidence bands is presented in fig. 5.

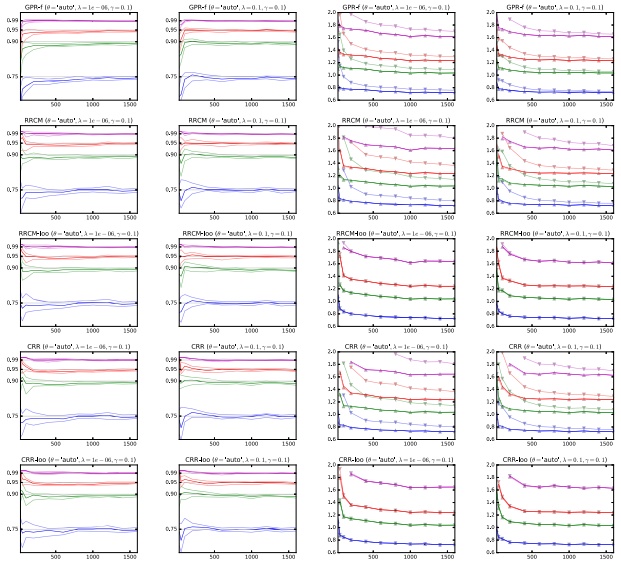
In the non-Gaussian setting the GPR confidence intervals are not consistently valid, as is evident from the coverage rate dynamics for $\lambda = 10^{-1}$. At the same time conformal procedures show no significant departures from claimed validity (results for other measures and residuals were qualitatively similar). The main conclusion is that in the negligible noise case the conformal confidence intervals for the KRR with the Gaussian kernel perform reasonably well both in terms of validity in a non-Gaussian setting, and efficiency in fully Gaussian setting.

The performance of conformal confidence regions in the noisy setting ($\gamma = 10^{-1}$) is qualitatively similar to the negligible noise case, except that the bands are wider due to higher observation noise. We report the findings for the MLE of θ only, but the results for conformal regions with fixed θ are qualitatively similar. In fig. 6 the conformal confidence regions provide the specified level of validity regardless of the parameter λ of the non-conformity measure. As expected, the Bayesian confidence predictions uphold their theoretical guarantees.

In the non-Gaussian setting with noise-to-signal ratio $\gamma = 10^{-1}$, fig. 7, all experiments yielded results similar to the negligible noise case: the conformal confidence sets are asymptotically valid, whereas the Bayesian intervals are not.

B. Results: 2-d

In this section we conduct experiments in the 2-d setting $\mathcal{X} = [-1,1]^2$, and the experimental steps are similar to IV-A The typical sample realisations of the studied $2-d$ functions f are depicted in fig. 8.



(a) (b) (c) (d)

Figure 6: Coverage rate and region size dynamics in the noisy fully Gaussian case with $\gamma = 10^{-1}$ for different $\theta = \hat{\theta}_{\text{ML}}$ and $\lambda = 10^{-6}$ (a), and $\lambda = 10^{-1}$ (b). Rows from *top* to *bottom*: "GPR-f", "RRCM", "RRCM-loo", "CRR", and "CRR-loo".

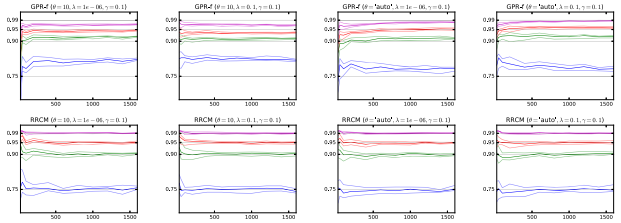


Figure 7: Coverage dynamics for the "Heaviside" ($\lambda = 10^{-1}$).

Table I shows the error rates ($y^* \notin \mathcal{B}(x^*)$) of the Bayesian confidence intervals on the fixed test sample. Columns 1 and 4 show that the regions are approximately valid when the kernel and noise hyper-parameters are known. The Bayesian intervals are more conservative for the case of low noise ($\gamma = 10^{-6}$) and high regularization $\lambda = 10^{-1}$. However, the validity of the GPR confidence intervals is sensitive to misspecification of kernel precision θ.

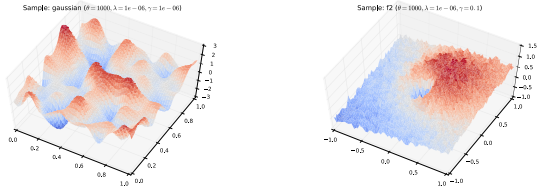


Figure 8: A sample path of a 2-d Gaussian process(*left* $\gamma = 10^{-6}$) and a non-Gaussian function "f2" (*right* $\gamma = 10^{-1}$).

Table I: The empirical error rate (%) of the GPR confidence interval for a simulated 2-d Gaussian process with train size $n = 1500$.

	γ	10^{-6}		10^{-1}	
	λ	10^{-6}	10^{-1}	10^{-6}	10^{-1}
θ	$\alpha(\%)$				
10^2	1	0.9	0.1	4.1	0.7
	5	4.5	0.5	12.0	4.4
	10	9.2	1.0	19.2	9.3
	25	23.8	3.3	36.1	24.5
$\hat{\theta}_{\mathrm{ML}}$	1	0.8	0.1	2.3	0.8
	5	4.4	0.5	4.9	4.5
	10	9.0	0.9	8.0	9.4
	25	23.6	2.3	18.9	24.7
10^1	1	1.6	0.6	1.0	0.6
	5	5.3	4.2	5.2	3.7
	10	9.6	8.6	10.4	8.2
	25	22.5	23.4	26.3	22.6
10^3	1	0.4	0.4	9.1	1.1
	5	1.2	1.2	13.1	4.9
	10	2.0	2.0	16.1	9.6
	25	4.7	4.5	24.0	24.0

Table II: The maximal absolute deviation $\mathtt{MAD}(\Gamma, A; \Theta)$ (%) of the empirical error rate from the theoretical significance level of conformal confidence regions for simulated 2-d Gaussian process for $n = 1500$.

	γ	10^{-6}		10^{-1}	
	λ	10^{-6}	10^{-1}	10^{-6}	10^{-1}
type	θ				
RRCM	10^1	0.3	0.2	0.8	0.4
	10^2	1.7	1.3	1.1	0.7
	10^3	1.1	2.6	1.2	0.5
	$\hat{\theta}_{\mathrm{ML}}$	1.7	2.2	0.1	0.5
RRCM-loo	10^1	1.3	0.3	2.0	0.4
	10^2	2.4	1.7	2.0	0.4
	10^3	2.9	2.6	0.1	0.6
	$\hat{\theta}_{\mathrm{ML}}$	2.6	2.6	0.8	0.6
CRR	10^1	0.3	0.1	0.8	0.3
	10^2	1.6	1.2	1.2	0.8
	10^3	0.8	2.2	1.3	0.5
	$\hat{\theta}_{\mathrm{ML}}$	1.8	2.1	0.1	0.4
CRR-loo	10^1	1.2	0.3	2.0	0.2
	10^2	2.4	1.7	2.1	0.5
	10^3	2.6	2.4	0.3	0.5
	$\hat{\theta}_{\mathrm{ML}}$	2.6	2.4	0.8	0.6

In contrast to the GPR confidence intervals, the conformal regions are insensitive to misspecification as shown in table II, where we report the maximal absolute deviation of the interval error rate from the specified rate α across all studied significance levels (21).

$$\mathtt{MAD}(\Gamma, A; \Theta) = \max_{\alpha \in A} \left| m^{-1} \#\{ j : y_j^* \notin \Gamma_n^\alpha(X_j^*; \Theta) \} - \alpha \right|, \quad (21)$$

where Θ is the vector of hyper-parameters of the experiment revealed to the conformal procedure, $A = \{1\%, 5\%, 10\%, 25\%\}$, $(X_j^*, y_j^*)_{j=1}^m$ is the test sample.

Typical profile of the test function used in non-Gaussian experiment is plotted in fig. 8. Performance of conformal regions in the non-Gaussian experiments is summarized in table IV. Overall, the error rates do not stray too far from the stated levels, and conformal regions are weakly sensitive to the KRR hyper-parameters. In contrast, table III shows that the empirical error rate of Bayesian confidence intervals depends more prominently on the values of the precision parameter. The MLE θ produces conservatively valid confidence intervals, and for $\gamma = 10^{-1}$ the error rate becomes closer to the specified significance level.

Table III: The empirical error rate (%) of the GPR confidence interval for the "f2" test function ($n = 1500$).

	γ	10^{-6}		10^{-1}	
	λ	10^{-6}	10^{-1}	10^{-6}	10^{-1}
θ	$\alpha(\%)$				
10^1	1	2.3	1.9	2.2	1.1
	5	3.2	3.0	7.9	4.9
	10	4.0	3.7	13.8	9.7
	25	5.8	5.6	29.9	24.3
10^2	1	0.3	0.0	19.1	1.3
	5	0.6	0.1	28.6	5.9
	10	0.9	0.2	35.0	11.1
	25	2.6	1.2	48.3	26.5
10^3	1	0.1	0.1	2.4	0.1
	5	1.9	2.1	4.0	1.9
	10	4.1	4.5	6.0	5.0
	25	12.9	13.4	13.9	16.4
$\hat{\theta}_{\mathrm{ML}}$	1	3.4	0.6	2.0	1.1
	5	4.4	1.1	3.9	5.1
	10	5.2	1.4	6.9	10.0
	25	7.1	2.4	18.1	24.9

Table IV: The maximal absolute deviation $\mathtt{MAD}(\Gamma, A; \Theta)$ (%) of the empirical error rate from the theoretical significance level of conformal confidence regions for the "f2" test function ($n = 1500$).

	γ	10^{-6}		10^{-1}	
	λ	10^{-6}	10^{-1}	10^{-6}	10^{-1}
type	θ				
CRR	10^1	1.0	1.3	1.2	0.2
	10^2	0.7	2.7	1.4	0.6
	10^3	0.3	1.1	1.0	0.1
	$\hat{\theta}_{\mathrm{ML}}$	1.4	2.2	0.6	0.5
CRR-loo	10^1	0.8	1.4	1.3	0.2
	10^2	3.0	3.2	1.9	0.5
	10^3	1.0	1.0	0.2	0.2
	$\hat{\theta}_{\mathrm{ML}}$	2.6	2.6	0.5	0.2
RRCM	10^1	0.9	1.4	1.2	0.2
	10^2	0.7	2.6	1.3	0.6
	10^3	0.4	0.5	0.9	0.3
	$\hat{\theta}_{\mathrm{ML}}$	1.3	2.3	0.8	0.4
RRCM-loo	10^1	0.8	1.6	1.2	0.3
	10^2	3.0	3.2	2.0	0.6
	10^3	0.9	0.6	0.2	0.2
	$\hat{\theta}_{\mathrm{ML}}$	2.6	2.6	0.7	0.1

V. Conclusion

Experiments in sec. IV provide evidence suggesting that conformal procedures are insensitive to the choice of the core NCM and are asymptotically equivalent in terms of coverage and efficiency, despite being applied in the off-line batch learning setting. Furthermore, the results indicate that both Bayesian and conformal confidence intervals possess asymptotic validity guarantees when the Gaussian assumptions hold, and that the conformal procedure yields asymptotically efficient regions.

Further research shall focus on establishing theoretical foundations for the obtained experimental evidence for the conformalized KRR with Gaussian kernel, isolating and studying cases when the guarantees are upheld, and generalizing the efficiency result in [23].

Acknowledgments

The research, presented in Section IV of this paper, was supported by the RFBR grants **16-01-00576 A** and **16-29-09649 ofi_m**; the research, presented in other sections, was conducted in IITP RAS and supported solely by the Russian Science Foundation grant (project **14-50-00150**).

References

[1] S. Alestra, E. Burnaev, C. Bordry, C. Brand, P. Erofeev, A. Papanov, and C. Silveira-Freixo, "Application of rare event anticipation techniques to aircraft health management," in *Mechanical and Aerospace Engineering V*, vol. 1016. Trans Tech Publications, 11 2014, pp. 413–417.

[2] E. Burnaev, P. Erofeev, and D. Smolyakov, "Model selection for anomaly detection," in *Proc. SPIE*, vol. 9875, 2015.

[3] A. Artemov, E. Burnaev, and A. Lokot, "Nonparametric decomposition of quasi-periodic time series for change-point detection," in *Proc. SPIE*, vol. 9875, 2015.

[4] C. C. Aggarwal and S. Y. Philip, "Outlier detection with uncertain data." in *Proc. of the SIAM International Conference on Data Mining*. SIAM, 2008, pp. 483–493.

[5] D. W. Scott, "Kernel density estimators," *Multivariate Density Estimation: Theory, Practice, and Visualization*, pp. 137–217, 2015.

[6] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proc. of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ser. ICPR '04. IEEE Computer Society, 2004, pp. 430–433.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, may 2000.

[8] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Loop: Local outlier probabilities," in *Proc. of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. ACM, 2009, pp. 1649–1652.

[9] M. F. Augusteijn and B. A. Folkert, "Neural network classification and novelty detection," *International Journal of Remote Sensing*, vol. 23, no. 14, pp. 2891–2902, 2002.

[10] S. Hawkins, H. He, G. J. Williams, and R. A. Baxter, "Outlier detection using replicator neural networks," in *Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings*, ser. DaWaK 2000. Springer Berlin Heidelberg, 2002, pp. 170–180.

[11] H. Hoffmann, "Kernel pca for novelty detection," *Pattern Recogn.*, vol. 40, no. 3, pp. 863–874, Mar. 2007.

[12] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[13] A. Bernstein, A. Kuleshov, and Y. Yanovich, *Manifold Learning in Regression Tasks*. Springer International Publishing, 2015, pp. 414–423.

[14] A. Kuleshov and A. Bernstein, "Extended regression on manifolds estimation," in *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proc.* Springer International Publishing, 2016, pp. 208–228.

[15] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[16] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. University Press Group Limited, 2006.

[17] E. V. Burnaev, A. A. Zaytsev, and V. G. Spokoiny, "The bernstein-von mises theorem for regression based on gaussian processes," *Russian Math. Surveys*, vol. 68, no. 5, p. 954, 2013.

[18] A. A. Zaitsev, E. V. Burnaev, and V. G. Spokoiny, "Properties of the posterior distribution of a regression model based on gaussian random fields," *Automation and Remote Control*, vol. 74, no. 10, pp. 1645–1655, 2013.

[19] M. Belyaev, E. Burnaev, and Y. Kapushev, "Gaussian process regression for structured data sets," in *Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings*. Springer International Publishing, 2015, pp. 106–115.

[20] M. Belyaev, E. Burnaev, and Y. Kapushev, "Computationally efficient algorithm for gaussian process regression in case of structured samples," *Computational Mathematics and Mathematical Physics*, vol. 56, no. 4, pp. 499–513, 2016.

[21] E. V. Burnaev, M. E. Panov, and A. A. Zaytsev, "Regression on the basis of nonstationary gaussian processes with bayesian regularization," *Journal of Communications Technology and Electronics*, vol. 61, no. 6, pp. 661–671, 2016.

[22] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[23] E. Burnaev and V. Vovk, "Efficiency of conformalized ridge regression," in *Proc. of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, 2014, pp. 605–622.