Nonparametric confidence bands for supervised learning

Noémi Takács

Supervisor: Ambrus Tamás, SZTAKI, ELTE

1 Introduction

In statistics binary classification is a fundamental problem. We observe a set of i.i.d. input-output pairs $\{(X_i, Y_i)\}$ in $\mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^d$ and the outputs are from two categories $\{-1, 1\}$. A typical task is to estimate the regression function $f_*(x) \doteq E[Y|X = x]$, from which we can calculate for a given input value, the probability that the output is from one of the categories:

$$\mathbb{P}(Y = 1|X = x) = \frac{f_*(x) + 1}{2}.$$

In general it is important to determine how close the estimate is to the reality. Constructing confidence regions are a usual way to evaluate the reliability. Obviously we want confidence regions with the best possible properties, such as containing the objective function with at least a given probability and in addition they are as small as possible. Also the constructing methods are determined by several properties e.g. how many and how strong assumptions they require.

In this project I considered the problem described using a certain algorithm, called the Sign-Perturbed Sums (SPS) [1] and its variants to create confidence intervals and regions for regression and classification [4] problems. Then I dealt with confidence bands built around the estimation of the regression function, again also in relation to regression [2] and classification.

1.1 Previous works

In the first semester I empirically analysed confidence intervals for mean estimation problems. I compared the performance of the SPS algorithm, which constructs non-asymptotic, distribution-free and exact confidence regions, to asymptotic methods. I also examined the simplest "classification problem" in which case there are no explanatory variables.

In the second part I made further steps in the investigation of binary classification and considered the problem with one explanatory variable. My aim was to estimate the regression function and construct confidence regions around a point estimator. For this I used the generalization of the SPS method. The algorithm was also demonstrated through simulations on synthetic and real data (e.g. Figure 1). I got promising results, but I had to choose a parameterized function family in which I approximated the real regression function.

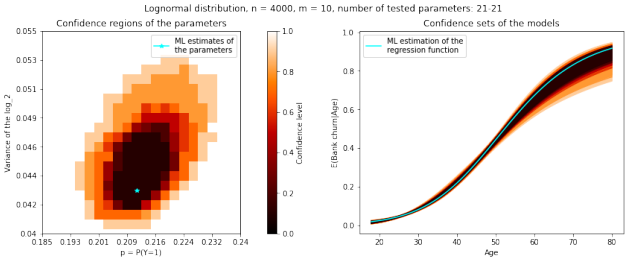

Figure 1: Application on real data: predicting bank churn using lognormal distribution family

Resampling framework

The SPS method and its modifications, are based on a resampling framework. The main idea is to generate $m - 1$ alternative outputs for the original inputs based on the conditional distribution defined by a parameter θ:

$$\mathbb{P}_\theta(Y = 1|X = x) = \frac{1 - f_\theta(x)}{2}.$$

Let $\mathcal{D}_0 = \{(X_j, Y_j)\}_{j=1}^n$ denote the original sample, then we construct the i-th alternative sample by

$$\mathcal{D}_i(\theta) \doteq \{(X_j, Y_{i,j}(\theta))\}_{j=1}^n,$$

where $Y_{i,j}(\theta)$ is generated from $\mathbb{P}_\theta(Y = 1|X_j)$. Here we have two observations:

1. If $\theta = \theta^*$, then $\mathcal{D}_0$ and $\mathcal{D}_i(\theta^*)$ comes from the same distribution.

2. If $\theta \neq \theta^*$, then the distribution of $\mathcal{D}_i(\theta)$ differs from that $\mathcal{D}_0$.

The significance of the difference can be detected with a statistical test, considering the following hypotheses:

$$H_0 : f_* = f_\theta$$
$$H_1 : f_* \neq f_\theta$$

1.2 Current study

In the third part of the project for constructing confidence bands and intervals I used nonparametric methods. The aim was to estimate the regression function in reproducing kernel Hilbert spaces and built around the estimation confidence bands.

First I present the necessary theoretical foundations: the construction of reproducing kernel Hilbert spaces, especially Paley-Wiener spaces and fitting via kernel ridge regression. I also made some plots to show how the kernel ridge estimation works for different regularization parameters, where we can see that with an inappropriate value the estimate will not be accurate enough or it will overfit the noise.

In the next section I introduce a method for constructing nonparametric, non-asymptotic and distribution-free confidence bands [2] for regression problems. I discuss this task in two parts: the case without noise, i.e. when the regression function is observed exactly, and the case with noisy observations. I also implemented the algorithms and demonstrated the method on synthetic examples.

Finally I describe an idea to reformulate the previous approach to binary classification. It has not yet been perfected, but it could be promising. The method would provide theoretical guarantees, but it is computationally demanding and requires further investigation to be applied in practice.

2 Preliminaries

Subsections 2.1 and 2.3 are based on [5].

2.1 Reproducing kernel Hilbert spaces (RKHS)

Let $\mathbb{V}$ be a vector space with an inner product. Since every inner product induces a norm: $\|f\|_\mathbb{V} := \sqrt{\langle f, f\rangle_\mathbb{V}}$, we can define the Cauchy sequences on this space in the usual way.

Definition 1 *A Hilbert space $\mathcal{H}$ is a complete inner product space $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$, i.e., every Cauchy sequence $(f_n)_{n=1}^\infty$ in $\mathcal{H}$ converges to some element $f^* \in \mathcal{H}$.*

Definition 2 *A linear functional on a Hilbert space is a mapping $L : \mathcal{H} \to \mathbb{R}$ that is linear, meaning that $L(f + \alpha g) = L(f) + \alpha L(g)$ for all $f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$. A linear functional is said to be bounded if there exists some $M < \infty$ such that $|L(f)| \leq M\|f\|_\mathcal{H}$ for all $f \in \mathcal{H}$.*

Theorem 1 (Riesz representation) *Let L be a bounded linear functional on a Hilbert space. Then there exists a unique $g \in \mathcal{H}$ such that $L(f) = \langle f, g\rangle_\mathcal{H}$ for all $f \in \mathcal{H}$.*

The reproducing kernel Hilbert spaces are spaces of real functions on a domain $\mathcal{X}$.

Definition 3 (PD kernel function) *A symmetric bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{x_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ Gram matrix with elements $K_{ij} := \mathcal{K}(x_i, x_j)$ is positive semidefinite.*

We say that a $\mathcal{K}$ kernel has the reproducing property for the $\mathcal{H}$ Hilbert space, if for any $x \in \mathcal{X}$, function $\mathcal{K}(\cdot, x)$ belongs to $\mathcal{H}$, and it satisfies

$$\langle f, \mathcal{K}(\cdot, x)\rangle_\mathcal{H} = f(x) \quad \forall f \in \mathcal{H}.$$

In particular:

$$\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, z)\rangle_\mathcal{H} = \mathcal{K}(x, z) \quad \text{for all } x, z \in \mathcal{X}.$$

Theorem 2 *Given any positive definite kernel function $\mathcal{K}$, there is a unique Hilbert space $\mathcal{H}$ in which the kernel satisfies the reproducing property. It is called the reproducing kernel Hilbert space associated with $\mathcal{K}$.*

Let $\tilde{\mathcal{H}}$ be a set of functions including every

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x_i), \tag{1}$$

where $n \in \mathbb{N}^+$, $\{x_i\}_{i=1}^n \subset \mathcal{X}$ is a set of points and $\alpha \in \mathbb{R}^n$ is a weight vector. This $\tilde{\mathcal{H}}$ forms a vector space. Let us define the inner product of two such functions as

$$\langle f, \bar{f} \rangle_{\mathcal{H}} := \sum_{i=1}^{n} \sum_{j=1}^{\bar{n}} \alpha_i \bar{\alpha}_j \mathcal{K}(x_i, \bar{x}_j).$$

By construction, this inner product satisfies the kernel reproducing property. Finally let $(f_n)_{n=1}^{\infty}$ be a Cauchy sequence in $\tilde{\mathcal{H}}$ and $f(x) \doteq \lim_{n\to\infty} f_n(x)$. If we extend $\tilde{\mathcal{H}}$ with all such $f(x)$, then we got a reproducing kernel Hilbert space, denoted by $\mathcal{H}$, due to theorem 2, where $\|f\|_{\mathcal{H}} \doteq \lim_{n\to\infty} \|f_n\|_{\tilde{\mathcal{H}}}$.

2.2 Paley-Wiener Spaces

Paley-Wiener space [2] consists of band-limited $f \in \mathcal{L}^2(\mathbb{R}, \lambda)$ functions, where λ is the Lebesgue measure, such that the support of the Fourier transform of f is included in $[-\eta, \eta]$, where $\eta > 0$. We can use the $\mathcal{L}^2$ inner product, because it is a subspace of the $\mathcal{L}^2$ space:

$$\langle f, g \rangle_{\mathcal{H}} \doteq \int_{\mathbb{R}} f(x) g(x) \, d\lambda(x)$$

This is an RKHS, and its reproducing kernel for $x \neq z \in \mathbb{R}$ is:

$$k(x, z) \doteq \frac{\sin(\eta(x - z))}{\pi(x - z)} \quad \text{and} \quad k(x, x) \doteq \frac{\eta}{\pi}.$$

2.3 Kernel ridge regression (KRR)

RKHSs are useful for solving classic statistical problems such as regression. In this problem we observe n noisy samples of $\{(x_i, y_i)\}_{i=1}^n$ input-output pairs. Suppose that $y_i = f^*(x_i) + w_i$ for $i \in [n]$, where $f^* : \mathcal{X} \to \mathbb{R}$ is an unknown function and w_i is the i-th measurement noise. We search for f^* in the finite form of 1.

If $w_i = 0 \ \forall i$ so there is no noise, the task is to find $\hat{f}$ which interpolates the observations. However there could be infinitely many functions in $\mathcal{H}$ such that $\hat{f}(x_i) = y_i \ \forall i$. We choose the one which has minimal RKHS norm, because this is the smoothest $\hat{f}$ of all possible. This leads to the optimization problem:

$$\arg\min_{f \in \mathcal{H}} \ \|f\|_{\mathcal{H}}, \quad \text{s.t. } f(x_i) = y_i \ \forall i. \tag{2}$$

In the presence of noise $\hat{f}$ should not fit perfectly to the data points, therefore we should introduce some trade-off between the fit and the Hilbert norm. Hence we only require that the differences between the observed and the fitted values be small. The modification of problem (2) is

$$\arg\min_{f \in \mathcal{H}} \ \|f\|_{\mathcal{H}}, \quad \text{s.t. } \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \delta^2,$$

where $\delta > 0$ is a tolerance parameter. Alternatively, we got the same f by solving

$$\min_{f \in \mathcal{H}} \ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2, \quad \text{s.t. } \|f\|_{\mathcal{H}} \leq R,$$

where $R > 0$ is an appropriately chosen radius. Both of these problems are convex and can be reformulated (by the Lagrangian duality) as

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{H}}^2 \right\}, \tag{3}$$

where $\lambda_n \geq 0$ is a regularization parameter, a function of δ or R.

One has to choose λ_n wisely, because this parameter is responsible for the smoothness of the estimated function. If it is too large, then the shape of the estimate may not be similar enough to the original function, hence the error will be large. At the other extreme, when λ_n is too small, then f can be very wavy fitting the noise. In practice choosing the optimal regularization (hyper)parameter can be done by e.g. cross-validation.

Theorem 3 *For all $\lambda_n \geq 0$, the kernel ridge regression estimate can be written as*

$$\hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\alpha}_i \mathcal{K}(\cdot, x_i),$$

where the optimal weight vector $\hat{\alpha} \in \mathbb{R}^n$ is given by

$$\hat{\alpha} = (K + \lambda_n I_n)^{-1} \frac{y}{\sqrt{n}},$$

where K is the kernel matrix multiplied by $\frac{1}{n}$.

2.3.1 Examples

I implemented the KRR algorithm and demonstrated on some example for both regression and binary classification. For these I used the Gaussian kernel:

$$k(x, z) \doteq \exp\left(-\frac{1}{2\sigma}\|x - z\|_2^2\right),$$

with σ = the standard deviation of X. For both example I generated 50 obervations.

1. *Regression:* X has uniform distribution on the interval $[0; 50]$, while the values of Y were calculated as follows: $max\left\{\frac{-(x-15)^2}{25} + 50; \frac{-(x-35)^2}{25} + 50\right\}$ plus noise from standard normal distribution. I chose the regularization parameter λ to be 0.001. This task is challenging, because the real regression function is not in the RKHS in question, but we can see on figure [2] that the method gives promising result.

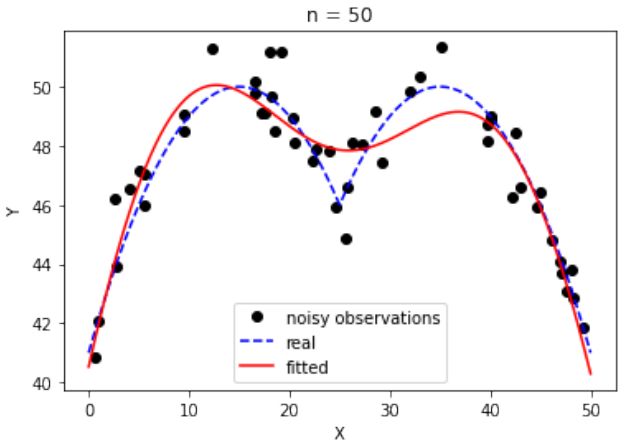

Figure 2: Fitted function by KRR for a continuous sample

2. *Binary classification:* As in my previous semester's work, I generated the points with Laplace distributions.

- $P(Y = 1) = P(Y = -1) = 0.5$,
- $f(X|Y = 1) =$ Laplace(location = 2, scale = 1),
- $f(X|Y = -1) =$ Laplace(location = -2, scale = 1).

Here I set λ to 0.05. The results are shown on figure [3].

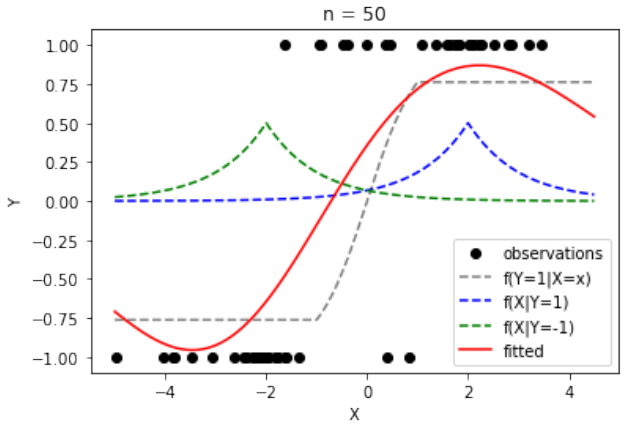

Figure 3: Fitted function by KRR for a binary sample

As I have already mentioned it is important to choose λ appropriately. In classification, the estimated function should be between -1 and 1. This can help us to fine-tune λ. Figures 4 and 5 show the effect of the parameter on the function.

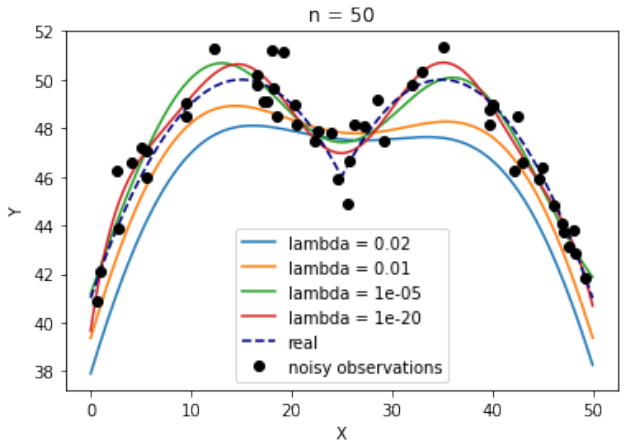

Figure 4: Regression estimate $\hat{f}$ for different regularization parameters.

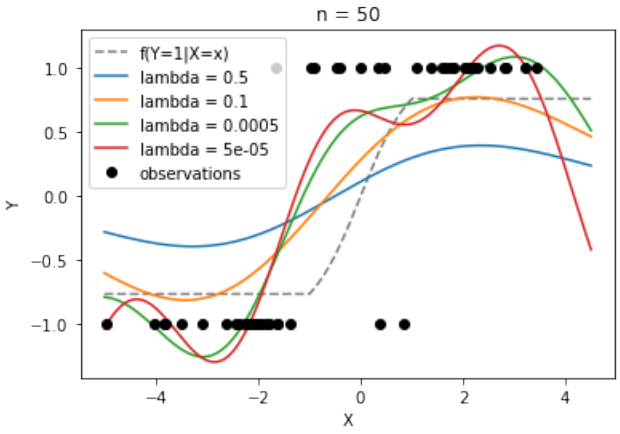

Figure 5: Regression function estimates $\hat{f}$ for classification for different regularization parameters.

3 Nonparametric confidence bands for regression problems

The purpose of this section is to construct confidence bands for the regression function based on [2]. In this report, I apply a nonparametric method for this problem, using Paley-Wiener kernel, i.e. compared to the previous semester, there is no need to define a parameterized model class. Formally the task is to find a function $I(x) = (I_1(x), I_2(x)) : \mathcal{D} \to \mathbb{R} \times \mathbb{R}$ such that

$$\nu(I) \doteq \mathbb{P}(\forall x \in \mathcal{D} : I_1(x) \leq f_*(x) \leq I_2(x)) \geq 1 - \alpha, \quad (4)$$

where $\mathcal{D}$ is the support of the input distribution, $\alpha \in (0,1)$ is a user-chosen risk probability and $\nu(I)$ is the reliability of the confidence band. With the following notation:

$$\mathcal{I} \doteq \{(x,y) \in \mathcal{D} \times \mathbb{R} : y \in [I_1(x), I_2(x)]\},$$

we can say that $\nu(I) = \mathbb{P}(\text{graph}_{\mathcal{D}}(f_*)) \subseteq \mathcal{I}$, where $\text{graph}_{\mathcal{D}}(f_*) \doteq \{(x, f_*(x)) : x \in \mathcal{D}\}$.

To build distribution-free, non-asymptotic and nonparametric confidence bands, the method requires the following assumptions:

(a1) The given input-output pairs $(x_1, y_1) \dots (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$, is an i.i.d. sample, such that $\mathbb{E}\left[y_k^2\right] < \infty \; \forall k \in [n]$. – Being i.i.d. is a standard assumption, square-integrability is for estimating the $\mathcal{L}^2$ norm of f_*.

(a2) The noise term, $\varepsilon_k \doteq y_k - f_*(x_k) \; \forall k \in [n]$ has a symmetric probability distribution about zero. – This is also a mild requirement.

(a3) The inputs $\{x_k\}$ have uniform distribution on $[0,1]$. – This is the strongest assumption. It can be relaxed to any known input distribution with a known strictly monotone increasing and continuous cumulative distribution function F, because then $x'_k \doteq F(x_k) \sim U(0,1)$.

(a4) f_* is included in a Paley-Wiener space; $\forall x \in [0,1] : |f_*(x)| \leq 1$ and f_* satisfies:

$$\int_{\mathbb{R}} f_*^2(x) \mathbb{I}(x \notin [0,1]) \, d\lambda(x) \;\leq\; \delta_0,$$

where $\mathbb{I}$ denotes the indicator function and δ_0 is a universal constant. – This assumption is needed to restrict the model class and to generalize effectively to unknown data points.

3.1 Noise-free case

First, I consider the problem without noise, i.e. I assume that $y_k = f_*(x_k) \; \forall k \in [n]$. The idea of the construction:

- assume that there exists a κ stochastic upper bound for the squared norm of the regression function,
- then include (x_0, y_0) in the confidence band if the function, which simultaneously interpolates this new point and the original input-output pairs, has a squared norm at most κ.

Since in the Paley-Wiener space the norm is the $\mathcal{L}^2$ norm and $y_k = f_*(x_k)$:

$$\frac{1}{n}\sum_{k=1}^{n} y_k^2 = \frac{1}{n}\sum_{k=1}^{n} f_*^2(x_k) \approx \mathbb{E}\left[f_*^2(X)\right] \approx \|f_*\|_2^2 = \|f_*\|_{\mathcal{H}}^2.$$

Lemma 1 *Assuming (a1), (a3), (a4) and that $y_k = f_*(x_k) \; \forall k \in [n]$, the following choice of κ:*

$$\kappa \doteq \frac{1}{n}\sum_{k=1}^{n} y_k^2 + \sqrt{\frac{\ln(\alpha)}{-2n}} + \delta_0$$

satisfies for any given $\alpha \in (0,1)$:

$$\mathbb{P}\left(\|f_*\|_{\mathcal{H}}^2 \leq \kappa\right) \geq 1 - \alpha.$$

Suppose that $x_0 \neq x_k \ \forall k \in [n]$. Now the aim is to find $[I_1(x_0), I_2(x_0)]$ according to 4. Let K_0 denote the Gram matrix of the observations extended with x_0. Then the minimum norm interpolation of $(x, \tilde{y}) \doteq (x_k, y_k)_{k=0}^n \ \forall y_0$:

$$\tilde{f}(x) = \sum_{k=0}^{n} \tilde{\alpha} k(x, x_k), \quad \text{where } \tilde{\alpha} = K_0^{-1} \tilde{y},$$

and the squared norm:

$$\|\tilde{f}\|_{\mathcal{H}}^2 = \tilde{\alpha}^\top K_0 \tilde{\alpha} = \tilde{y}^\top K_0 \tilde{y}.$$

Finally we can calculate the minimum and maximum value of y_0, which satisfies the requirements, by solving separately the following optimization problems:

$$\begin{aligned} \min / \max \quad & y_0 \\ \text{s.t.} \quad & (y_0, y^\top) K_0^{-1} (y_0, y^\top)^\top \leq \kappa. \end{aligned} \tag{5}$$

This is a convex problem, but there is also an analytical solution, which is detailed in table 1.

1. Calculate $\kappa \doteq \frac{1}{n} \sum_{k=1}^n y_k^2 + \sqrt{\frac{\ln(\alpha)}{-2n}} + \delta_0$.
2. Create the extended Gram matrix:
$$K_0(i+1, j+1) \doteq k(x_i, x_j) \ \forall i, j \in [n].$$
3. Determine K_0^{-1} (exists, because K_0 is PSD) and its following partition:
$$K_0^{-1} = \begin{bmatrix} c & b^\top \\ b & A \end{bmatrix}.$$
4. Calculate the solutions $y_{min} \leq y_{max}$ of the quadratic equation $a_0 y_0^2 + b_0 y_0 + c_0$, where $a_0 \doteq c$, $b_0 \doteq 2b^\top y$, $c_0 \doteq y^\top A y - \kappa$.
5. Return $I_1(x_0) \doteq y_{min}$ and $I_2(x_0) \doteq y_{max}$. If there is no solution, then $I(x_0) \doteq \emptyset$.

Table 1: Pseudocode for the confidence interval in the noise-free case

3.1.1 Simulation

I implemented this algorithm and made a numerical demonstration, similar to the one in [2]. The regression function f_* from which I generated the observations had the form:

$$f_*(x) = \sum_{k=1}^{20} w_k (x, \bar{x}_k), \tag{6}$$

where $\{x_k\}_{k=1}^{20} \sim U(0,1)$ are random input points and $\{w_k\}_{k=1}^{20} \sim U(-1,1)$ are random weights. Finally I divided it with the maximum in absolute value from the $[0,1]$ interval to restrict the output values to $[-1,1]$. I set the other parameters as follows:

- $\eta = 30$ for the Paley-Wiener kernel,
- $\delta_0 = 0$,
- $\alpha = 0.5$ and 0.1,
- $n = 10$ observation.

The results: the real regression function, the observations and the two confidence bands are shown in Figure 6.

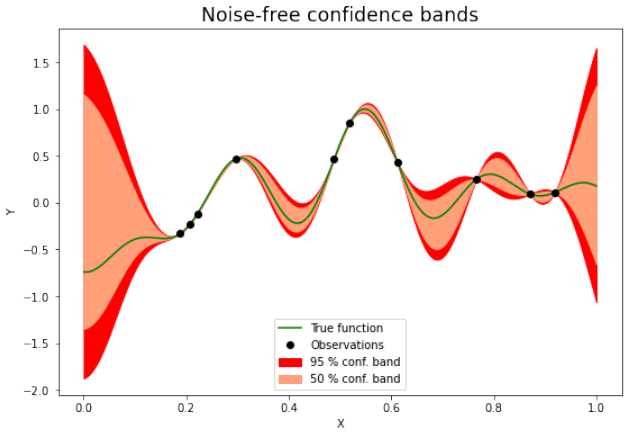

Figure 6: Confidence bands for a noise-free regression function

3.2 Noisy observations

In this subsection I deal with the noisy problem, so I assume that $y_k = f_*(x_k) + \varepsilon_k \ \forall k \in [n]$. The idea of the construction:

- build simultaneous confidence intervals for some observed points, and use these for bounding the norm,
- make confidence interval for an unobserved input, using the upper bound for the norm and the information, that the previously selected points are in the already calculated intervals with some probability.

Confidence intervals for observed points:

For building simultaneous confidence intervals for some selected points (let the number of it be d, $d \leq n$) I use kernel gradient perturbation (KGP) like in [2], which is an extension of the Sign-Perturbed Sums method (I used its variations in the previous semesters). The KGP method builds intervals for the RKHS coefficients around a kernel estimation, here around the kernel ridge regression. Problem 3 can be redefined as follows:

$$\min \quad \frac{1}{n}(y - K_1\theta)^\top W(y - K_1\theta) + \lambda\theta^\top K_2\theta, \quad (7)$$

where $K_1 \in \mathbb{R}^{n\times d}$ is the first d columns of K, $K_2 \in \mathbb{R}^{d\times d}$ is the first d row of K_1 and W is a diagonal matrix, which contains the given weights. This means that all observations are still used to calculate the error, but we look for $\tilde{\theta} \in \mathbb{R}^d$.

With the following notations,

$$\Phi = \begin{bmatrix} \frac{1}{\sqrt{n}} W^{\frac{1}{2}} K_1 \\ \sqrt{\lambda} K_2^{\frac{1}{2}} \end{bmatrix}, \quad v = \begin{bmatrix} \frac{1}{\sqrt{n}} W^{\frac{1}{2}} y \\ 0_d \end{bmatrix},$$

the objective of 7 can be reformulated as an ordinary least squares problem, $\|v - \Phi\theta\|^2$, of which solution is well known: $\hat{\theta} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top v$.

In this case the KGP confidence regions are the same as the ones produced by SPS. They are star convex around $\hat{\theta}$, and have ellipsoidal outer approximations with a given $1 - \beta \in (0, 1)$ confidence probability:

$$\hat{\Theta}_\beta \doteq \left\{\theta \in \mathbb{R}^d : (\theta - \hat{\theta})^\top \frac{1}{n}\Phi^\top\Phi(\theta - \hat{\theta}) \leq r\right\},$$

where r is the radius of the confidence ellipsoid. The computation of the radius can be done by semi-definite programming [1].

Let q, m integers, such that $\frac{q}{m} = \beta$ and introduce the following notations:

$$R_n \doteq \frac{1}{n}\Phi^\top\Phi,$$
$$\epsilon_k(\theta) \doteq v_k - \Phi_k^\top\theta,$$
$$S_0(\theta) \doteq R_n^{-\frac{1}{2}} \frac{1}{n}\sum_{k=1}^{n+d} \Phi_k\epsilon_k(\theta),$$
$$S_i(\theta) \doteq R_n^{-\frac{1}{2}} \frac{1}{n}\sum_{k=1}^{n+d} \alpha_{i,k}\Phi_k\epsilon_k(\theta),$$

where $\alpha_{i,k}$ is a random sign for $i \in [m-1]$ and $k \in [d]$, and $\alpha_{i,k} = 1$ for $i \in [m-1]$ and $k = d+1, \ldots, n$. Since $\|S_0(\theta)\|^2$ can be rewritten as $(\theta - \hat{\theta}_n)^\top R_n(\theta - \hat{\theta}_n)$, based on the SPS method r will be the q-th largest value of $\{\|S_i(\theta)\|^2\}_{i=1}^{m-1}$.

Expanding the expression $\|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2$ we got:

$$(\theta - \hat{\theta}_n)^\top R_n(\theta - \hat{\theta}_n) \leq$$
$$\left[\frac{1}{n}\sum_{k=1}^{n+d} \alpha_{i,k}\Phi_k(v_k - \Phi_k^\top\theta)\right]^\top R_n^{-1}\left[\frac{1}{n}\sum_{k=1}^{n+d} \alpha_{i,k}\Phi_k(v_k - \Phi_k^\top\theta)\right] =$$
$$= \theta^\top Q_i R_n^{-1} Q_i\theta - 2\theta^\top Q_i R_n^{-1}\psi_i + \psi_i^\top R_n^{-1}\psi_i,$$

where $Q_i \in \mathbb{R}^{d\times d}$ and $\psi_i \in \mathbb{R}^d$ are defined as

$$Q_i \doteq \frac{1}{n+d}\sum_{k=1}^{n+d} \alpha_{i,k}\Phi_k\Phi_k^\top,$$
$$\psi_i \doteq \frac{1}{n+d}\sum_{k=1}^{n+d} \alpha_{i,k}\Phi_k v_k.$$

Let z denote the quantity $R_n^{\frac{1}{2}T}$, then we can calculate $\max_{\theta:\|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2} \|S_i(\theta)\|^2$ by solving the following optimization problem:

$$\begin{aligned} \max \quad & \|z\|^2 \\ \text{s.t.} \quad & z^\top A_i z + 2z^\top b_i + c_i \leq 0, \end{aligned} \quad (8)$$

where A_i, b_i and c_i are as follows:

$$A_i \doteq I - R_n^{-\frac{1}{2}} Q_i R_n^{-1} Q_i R_n^{-\frac{1}{2}T},$$
$$b_i \doteq R_n^{-\frac{1}{2}} Q_i R_n^{-1}(\psi_i - Q_i\hat{\theta}_d),$$
$$c_i \doteq -\psi_i^\top R_n^{-1}\psi_i + 2\hat{\theta}_d^\top Q_i R_n^{-1}\psi_i - \hat{\theta}_d^\top Q_i R_n^{-1} Q_i\hat{\theta}_d.$$

In general problem 8 is not convex, but due to strong duality its solution is the same as the solution of its dual, which is convex:

$$\begin{aligned} \min \quad & \gamma \\ \text{s.t.} \quad & \lambda \geq 0, \\ & \begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^\top & \lambda c_i + \gamma \end{bmatrix} \succeq 0, \end{aligned} \quad (9)$$

where "$\succeq 0$" means positive definiteness.

Let γ_i^* denote the solution of 9. Then, choosing r as the q-th largest γ_i^*, it will be true that

$$\mathbb{P}(\tilde{\theta} \in \hat{\Theta}_\beta) \geq 1 - \frac{q}{m} = 1 - \beta.$$

Now we can give a lower and an upper bound for $f_*(x_k)$ $\forall k \in [d]$. Let $\varphi_k \doteq$

$(k(x_1, x_k), \ldots, k(x_d, x_k))^\top$, hence $f_*(x_k) = \varphi_k^\top \tilde{\theta}$. However, all we know is that $\hat{\Theta}_\beta$ includes $\tilde{\theta}$ with some user-chosen probability. Using this information we minimize and maximize $\varphi_k^\top \theta$, such that $\theta \in \hat{\Theta}_\beta$. We can calculate these values analytically. The solutions are

$$\nu_k \doteq \varphi_k \hat{\theta} - (\varphi_k^\top P \varphi_k)^{\frac{1}{2}},$$
$$\mu_k \doteq \varphi_k \hat{\theta} + (\varphi_k^\top P \varphi_k)^{\frac{1}{2}},$$

where $P = d \cdot r \cdot (\Phi^\top \Phi^\top)^{-1}$. Thus $\{[\nu_k, \mu_k]\}$ satisfies

$$\mathbb{P}(\forall k \in [d] : f_*(x_k) \in [\nu_k, \mu_k]) \geq 1 - \beta. \qquad (10)$$

Confidence intervals for unobserved points:

In the noise-free case Lemma 1 gave us an upper bound for the RKHS norm with probability at least $1 - \alpha$. Now we know that $f_*^2(x_k) \leq \max\{\nu_k^2, \mu_k^2\}$ $\forall k \in [d]$ with probability at least $1 - \beta$.

Lemma 2 *Assuming (a1), (a3), (a4) and the confidence intervals $[\nu_k, \mu_k]$ fulfill 10 $\forall k \in [d]$, then the following choice of τ:*

$$\tau \doteq \frac{1}{d} \sum_{i=1}^{d} \max\{\nu_k^2, \mu_k^2\} + \sqrt{\frac{ln(\alpha)}{-2d}} + \delta_0$$

satisfies for any given $\alpha, \beta \in (0, 1)$:

$$\mathbb{P}(\|f_*\|_{\mathcal{H}}^2 \leq \tau) \geq 1 - \alpha - \beta.$$

Suppose that $x_0 \neq x_k \forall k \in [d]$. Let K_0 denote the extended gram matrix:

$$K_0(i+1, j+1) \doteq k(x_i, x_j) \quad \text{for } i, j = 0, \ldots, d.$$

Now we have to solve separately similar problems to 5, but in this case we do not know the exact value of $f_*(x_k)$, $k \in [d]$, so they will also be variables from certain intervals:

$$\begin{aligned} &\min / \max\ z_0 \\ &\text{s.t. } (z_0, \ldots, z_d) K_0^{-1} (z_0, \ldots, z_d)^\top \qquad (11) \\ &\nu_1 \leq z_1 \leq \mu_1, \ldots, \nu_d \leq z_d \leq \mu_d. \end{aligned}$$

These are convex problems. If there is no solution, then $I(x_0) \doteq \emptyset$, otherwise $I_1(x_0) \doteq z_{min}$ and $I_2(x_0) \doteq z_{max}$.

In [3] some refinement for this algorithm are presented: one can e.g. relax the assumption on the noise term, by also allowing non-symmetric noises, introducing a more efficient norm estimating method and enhancing the construction of the confidence bands, by replacing the constraints with tighter ones.

3.2.1 Simulation

I implemented the whole algorithm and tested how it works in practice. For the convex optimization problems I used a Python package called CVXPY, which can even solve convex optimization problems including semi-definite programs.

The true regression function was the same as in the noise-free case (6). The noise term, ε had Laplace distribution with location, $\mu = 0$ and scale, $\lambda = 0.4$ parameters. The rest of the parameters were as follows:

- $\eta = 30$ for the Paley-Wiener kernel,
- $\delta_0 = 0$,
- $\alpha = \beta = 0.25$ and 0.05,
- $n = 100$ and $d = 20$,
- the regularization parameter $\lambda = 0.01$.

The results are summarized in Figure 7: the true regression function, the KRR estimate of the regression function, the observations, the selected points for bounding the RKHS norm and the two confidence bands are shown.

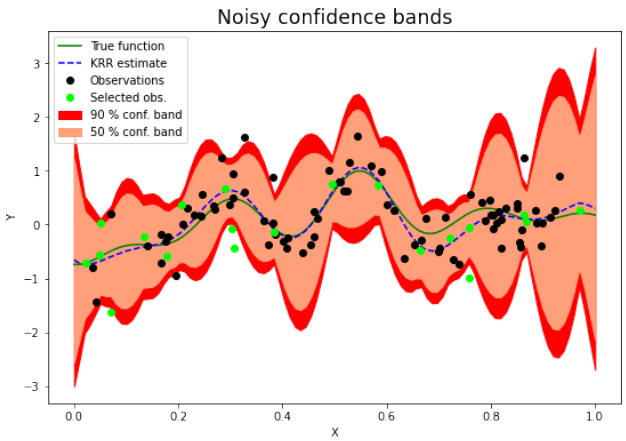

Figure 7: Confidence bands for noisy observations

4 Nonparametric confidence bands for binary classification

One goal of this semester was to reformulate the previous method (for noisy regression) to solve binary classification problems. In this case we cannot use exactly the same algorithm, because the noise term is not symmetric, so we cannot generate new samples in the same way as before. We have a concept for the modification, but how to put it into practice needs further research.

For simplicity, I present the idea for confidence level 50% (so only one resampling is required).

We can start the same way: choose d from the n observations, and as mentioned before, calculate K_1 and K_2. By solving problem 7, where $y \in \{-1, 1\}$, we got an estimate for the true θ, let it be denoted by $\hat{\theta}$. The new sample for the output is generated as follows:

$$\bar{y}(\theta) = \text{sign}(K_2\theta + U),$$

where $U \in \mathbb{R}^d$ has a uniform distribution on $[-1, 1]^d$. Suppose that $\bar{\theta}(\theta)$ is the estimate of θ from the new sample. With this notations, the quantities required for testing θ:

$$\begin{aligned} S_0(\theta) &\doteq \|f_\theta - \hat{f}_{KRR}(\mathcal{D}_0)\|^2_{\mathcal{H}} = \left\|\sum_{i=1}^{d} \theta_i k(\cdot, x_i)\right\|^2_{\mathcal{H}} \\ &= \theta^\top K_2\theta + \hat{\theta}^\top K_2\hat{\theta} - 2\theta^\top K_2\hat{\theta}, \\ S_1(\theta) &\doteq \|f_\theta - \hat{f}_{KRR}(\mathcal{D}_1(\theta))\|^2_{\mathcal{H}} \\ &= \theta^\top K_2\theta + \bar{\theta}^\top(\theta) K_2\bar{\theta}(\theta) - 2\theta^\top K_2\bar{\theta}(\theta). \end{aligned}$$

The decision whether including f_θ in the confidence band: if $S_1(\theta) > S_0(\theta)$ then accept θ. This expression can be expanded in the following way:

$$\hat{\theta}^\top K_2\hat{\theta} - 2\theta^\top K_2\hat{\theta} - \bar{\theta}^\top(\theta) K_2\bar{\theta}(\theta) + 2\theta^\top K_2\bar{\theta}(\theta) \leq 0.$$

Since we know $\hat{\theta}$ and K_2 let $a \doteq \hat{\theta}^\top K_2\hat{\theta}$ and $b \doteq 2K_2\hat{\theta}$. Furthermore $\forall i \in [d]$:

$$k_{x_i} \doteq (k(x_i, x_1), \ldots, k(x_i, x_d))^\top.$$

Now we can write up the optimization problem 8.

$$\begin{aligned} \max_\theta \ &= f_\theta(x_i) = \theta^\top k_{x_i} \\ \text{s.t.} \ &= a - \theta^\top b - \\ &\quad \text{sign}(K_2\theta + U)^\top \frac{1}{\sqrt{n}} (K_2 + \lambda I)^{-1} K_2 (K_2 + \lambda I)^{-1} \cdot \\ &\quad \frac{1}{\sqrt{n}} \text{sign}(K_2\theta + U) + \\ &\quad 2\theta^\top K_2 (K_2 + \lambda I)^{-1} \frac{1}{\sqrt{n}} \text{sign}(K_2\theta + U) \leq 0 \end{aligned} \tag{12}$$

It is difficult to solve because of the term $\text{sign}(K_2\theta + U)$, but we can consider all of the possible value of this and solve the problems separately. In practice, this can really increase the running time, since in this case there are 2^d tasks to solve. Let us introduce the notation

$$e \doteq (\pm 1, \ldots, \pm 1)^\top \in \mathbb{R}^d,$$

for a certain value of $\text{sign}(K_2\theta + U)$. Now we can rewrite the constraint of the problem 12, by replacing $\text{sign}(K_2\theta + U)$ with e, and adding that they are equal. This can be expressed as the product of $e^\top$ and $(K_2\theta + U)$ is non-negative, since their signs are the same. Hence the optimization problem is:

$$\begin{aligned} \max_\theta \ &= \theta^\top k_{x_i} \\ \text{s.t.} \ &= e_i(\theta^\top k_{x_i} + U_i) \geq 0, \quad \forall i \in [d] \quad \text{and} \\ &\quad a - \theta^\top b - \\ &\quad e^\top \frac{1}{\sqrt{n}} (K_2 + \lambda I)^{-1} K_2 (K_2 + \lambda I)^{-1} \frac{1}{\sqrt{n}} e + \\ &\quad 2\theta^\top K_2 (K_2 + \lambda I)^{-1} \frac{1}{\sqrt{n}} e \ \leq \ 0 \end{aligned} \tag{13}$$

Now these are easier problems, because they can be solved by linear programming.

The endpoints of 50 % confidence intervals for the observation $x_i, \ i = 1, \ldots, d$ are the minimum and the maximum value of the solutions of 13. In the case, when the minimum is lower than -1 and/or the maximum exceeds 1, we can define the interval endpoints as -1 and 1.

As I mentioned before there are some open questions about this algorithm. It works in theory, but its computational demand depends exponentially on the number of observations, which makes it difficult to apply in practice. Another question, is it possible to smooth the confidence bands, for

example by running the algorithm several times, always choosing different points to construct simultaneous confidence intervals, and finally averaging the endpoints of the intervals we got? In the future, it would be worth working on it further and finding ways to refine the current version.

5 Conclusions

In the 3 semesters of the project I tried to capture the uncertainty of different estimates with binary classification problems as its main objective. For constructing exact confidence intervals, regions and bands I used a non-asymptotic and distribution-free method (and its modifications) called the Sign-Perturbed Sums, which is based on a resampling framework.

In this report I presented nonparametric methods to estimate the regression function and build confidence bands, based on the theory of Paley-Wiener reproducing kernel Hilbert spaces. For estimating the regression function, I used kernel ridge regression, and then I constructed confidence bands around this estimation in noise-free and noisy regression. I also described an idea to reformulate this method for binary classification problems. The algorithms provides stochastic guarantees for small sample sizes, and was validated with numerical simulations in the regression problems.

References

[1] Csáji, B. Cs., Campi, M. C., & Weyer, E. (2015). Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models. *IEEE Transactions on Signal Processing*, *63*(1), 169–181. https://doi.org/10.1109/tsp.2014.2369000

[2] Csáji, B. Cs., & Horváth, B. (2022). Nonparametric, Nonasymptotic Confidence Bands With Paley-Wiener Kernels for Band-Limited Functions. *IEEE Control Systems Letters*, *6*, 3355–3360. https://doi.org/10.1109/lcsys.2022.3185143

[3] Csáji, B. Cs., & Horváth, B. (2023). Improving Kernel-Based Nonasymptotic Simultaneous Confidence Bands. *IFAC-PapersOnLine*, *56*(2), 10357–10362. https://doi.org/10.1016/j.ifacol.2023.10.1047

[4] Tamás, A., & Csáji, B. Cs. (2020). Sztochasztikus garanciák bináris klasszifikációhoz. *Alkalmazott Matematikai Lapok*, *37*, 365–379. https://doi.org/10.37070/AML.2020.37.2.16

[5] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.