Nonparametric confidence bands for supervised learning

Author:
Noémi Takács

Supervisor:
Ambrus Tamás
SZTAKI, ELTE

Math Project ELTE, January 2025

Table of contents

Topic of the project and the previous works

Binary classification and regression problems

→ estimate the regression function

→ construct confidence sets around the estimation

First semester:

- confidence intervals for mean estimates
- preparation for binary classification

Second semester:

- estimate the regression function (f_*) in binary classification
- construct parameterized confidence regions around the estimation

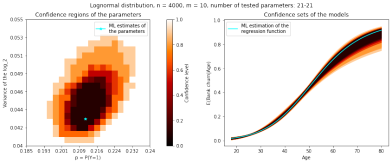

Figure: Predicting bank churn using lognormal distribution family

Current study

- Nonparametric methods
  - $\rightarrow$ reproducing kernel Hilbert spaces
  - $\rightarrow$ kernel ridge regression
- Confidence bands for regression problems
  - $\rightarrow$ noise-free case
  - $\rightarrow$ noisy case
- Confidence bands for binary classification
  - $\rightarrow$ an idea to reformulate the algorithm used for the regression with noisy observations

Sign-Perturbed Sums (SPS)

Main idea of the algorithm

- generate alternative outputs for the original inputs (perturb the residuals)
- compare the original $\mathcal{D}_0$ and the alternative samples $\{\mathcal{D}_i\}_{i=1}^{m-1}$ with a ranking function
- construct confidence set based on the rank of $\mathcal{D}_0$

Advantages:

- mild statistical assumptions
- distribution-free
- non-asymptotic
- exact confidence sets

Reproducing Kernel Hilbert Spaces (RKHS)

Let $\mathcal{H}$ be a Hilbert space $\to (\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

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 satisfies

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

Especially:

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

Theorem

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 $\mathcal{K}$.

Paley-Wiener spaces:

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

$\to$ we can use the $\mathcal{L}^2$ inner product

Kernel Ridge Regression (KRR)

$y_i = f^*(x_i) + \varepsilon_i$
We search for f^* in the finite form:

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

Noise-free case:
Interpolate the observations $\rightarrow$ infinitely many $\hat{f}$ $\rightarrow$ choose the one with minimal RKHS norm (smoothest) $\rightarrow$ solve the optimization problem:

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

Noisy case:
Trade-off between the fit and the Hilbert norm $\rightarrow$ solve the optimization problem:

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

where $\lambda_n \geq 0$ is a regularization parameter.

Examples for KRR and the choice of λ I.

Gaussian kernel:

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

Regression:

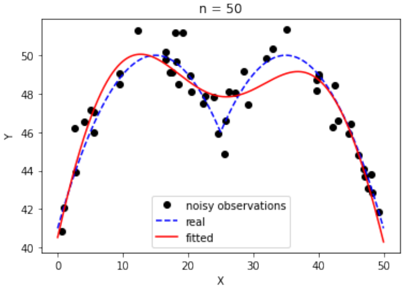

Figure: Fitting via KRR for continuous sample, $\lambda = 0.001$

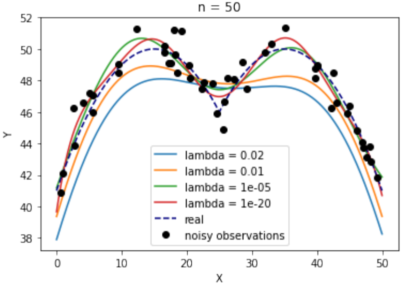

Figure: KRR estimates with different regularization parameters in regression

Examples for KRR and the choice of λ II.

Binary classification:

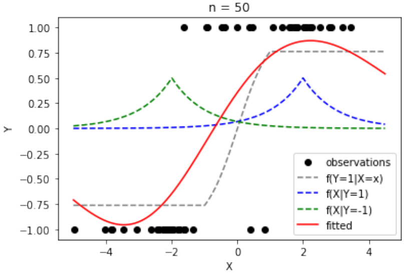

Figure: Fitting via KRR for binary sample, $\lambda = 0.05$

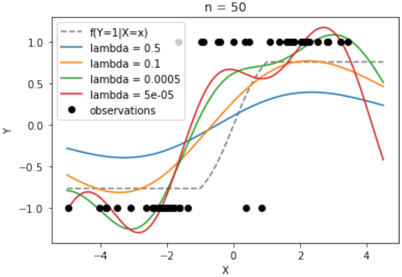

Figure: KRR estimates with different regularization parameters in binary classification

Nonparametric confidence bands for regression problem

Experiments based on article [2].

$\rightarrow$ Paley-Wiener kernel

The task: find a function $I(x) = (I_1(x), I_2(x)) : \mathcal{D} \to \mathbb{R} \times \mathbb{R}$

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

Assumptions:

- 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 $\mathbb{E}\left[y_k^2\right] < \infty \; \forall k \in [n]$.
- The noise term, $\varepsilon_k \doteq y_k - f_*(x_k) \; \forall k \in [n]$ has a symmetric probability distribution about zero.
- The inputs $\{x_k\}$ have uniform distribution on $[0, 1]$.
- 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.

Construction basics for noise-free regression

No noise $\rightarrow y_k = f_*(x_k)\ \forall k \in [n]$

Idea of the construction:

1. Assume that there exists a κ stochastic upper bound for the squared norm of the regression function.
   - $\rightarrow$ Since we can use the $\mathcal{L}^2$ norm:

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

   - $\rightarrow$ **Lemma:**

$$\text{with } \kappa \doteq \frac{1}{n}\sum_{k=1}^{n} y_k^2 + \sqrt{\frac{\ln(\alpha)}{-2n}} + \delta_0, \quad \mathbb{P}\left(\|f_*\|_{\mathcal{H}}^2 \leq \kappa\right) \geq 1-\alpha.$$

2. 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.
   - $\rightarrow$ Finally, there are 2 (convex) optimization problems to solve (which also have analytical solutions):

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

Simulation for noise-free regression

The true regression function:

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

divided by $\max_{x \in [0,1]} f_*(x)$, 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.

The other parameters:

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

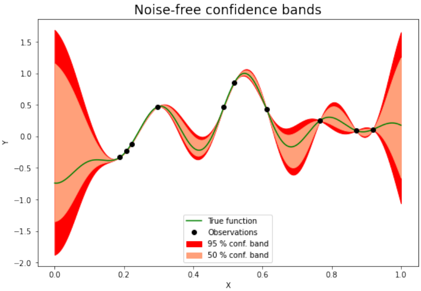

Figure: Confidence bands for a noise-free regression function

Construction basics for noisy regression

Noisy observations $\rightarrow y_k = f_*(x_k) + \varepsilon_k \; \forall k \in [n]$

Idea of the construction:

1. Build simultaneous confidence intervals for some observed points (select d, $d \leq n$), and use these for bounding the norm.
   - $\rightarrow$ with Kernel Gradient Perturbation (KGP) (extension of the SPS) build confidence intervals for the RKHS coefficients around the KRR estimation:

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

   - $\rightarrow$ **Lemma:**

$$\text{with } \tau \doteq \frac{1}{d} \sum_{i=1}^{d} \max\{\nu_k^2, \mu_k^2\} + \sqrt{\frac{ln(\alpha)}{-2d}} + \delta_0, \quad \mathbb{P}(\|f_*\|_{\mathcal{H}}^2 \leq \tau) \geq 1 - \alpha - \beta.$$

2. 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.
   - $\rightarrow$ The (convex) optimization problems:

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

Simulation for noisy regression

The true regression function is the same as before. The noise term:

$$\varepsilon \sim \textit{Laplace}(\textit{location} = 0, \textit{scale} = 0.4)$$

The other parameters:

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

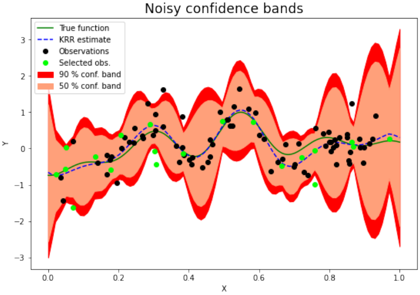

Figure: Confidence bands for a noisily observed regression function

Challenges in binary classification compared to regression

Binary observations $\rightarrow y_k \in \{-1, 1\}\ \forall k \in [n]$

If confidence interval endpoints are out of $[-1, 1] \rightarrow$ define them as -1 and/or 1.

Question:
The noise term is not symmetric $\rightarrow$ cannot generate new samples in the same way.

Idea for new sample generation:

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

where $U \sim U([-1, 1]^d)$.

Problem:
This term appears (multiple times) in the constraint of the optimization task to compute confidence intervals built around the observed points.

Solution idea

Idea to avoid the signum function:

Consider all possible values.

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

Replacing implies adding to the constraints:

$$e_i(\theta^\top k_{x_i} + U_i) \geq 0, \quad \forall i \in [d],$$

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

- $\rightarrow$ easier to solve, but
- $\rightarrow$ 2^d task (increasing exponentially with the number of observations)

Conclusions

The algorithm provides stochastic guarantees also for small sample sizes

$\rightarrow$ validated with numerical simulations in regression problems

Questions and improvement opportunities in binary classification:

- further investigation for practical applicability
- find a way to construct confidence bands with less computational demand
- make smoother bands, e.g. with averaging

References

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

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

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

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

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

Thank you for your attention!