Team: #3

Course Project
22/12/2023

Team Project on the course “Principles of Applied Statistics”

A Kernel Test of Goodness of Fit

Kamil **Garifullin** Viktoriia **Zinkovich** Maksim **Osipenko**

[1] **A Kernel Test of Goodness of Fit**. Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton

Problems
Problems
Problems

Problems

Motivation for the research

Problems
Problems
Problems
Problems

Problem Statement

Goal: if given a set of sample $\{Z_i\}_{i=1}^{n}$ with distributio $Z_i \sim q$, our interest is in whether **q matches** some reference or **target distribution p**

Gorham & Mackey's (2015) approach problems:

Complexity of the function class used (results from applying the Stein operator to the Sobolev space)

Unclear how to compute **p-values** or determine when to accept the null hypothesis

Methods
Methods
Methods

Methods

Theoretical methods used in the following work

Methods
Methods
Methods
Methods

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,\boxed{(T_p f)}(Z) - \mathbb{E}\,(T_p f)(X)$$

$$\boxed{T_p f} := \sum_{i=1}^{d} \left(\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} \right)$$

Stein operator
acting on $f \in \mathcal{F}^d$

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

It can be shown that

$$S_p(Z)^2 = \mathbb{E}h_p(Z, Z')$$

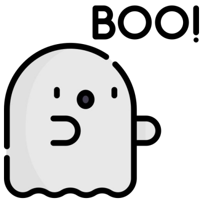
Simplify the equation!

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$



It can be shown that

$$S_p(Z)^2 = \mathbb{E}h_p(Z, Z')$$

$$\begin{aligned} h_p(x,y) :=& \nabla \log p(x)^\top \nabla \log p(y) k(x,y) + \nabla \log p(y)^\top \nabla_x k(x,y) \\ &+ \nabla \log p(x)^\top \nabla_y k(x,y) + \langle \nabla_x k(x,\cdot), \nabla_y k(\cdot,y) \rangle_{\mathcal{F}^d} \end{aligned}$$

Methods: Definitions

Goal: find the maximum discrepancy between **target distribution p** and **observed sample distribution q** in a RKHS (*Reproducing Kernel Hilbert Space*) $\mathcal{F}$

For that task – we define a **Stein discrepancy:**

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

It can be shown that

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z')$$

$$\begin{aligned} h_p(x,y) :=& \nabla \log p(x)^\top \nabla \log p(y) \boxed{k(x,y)} + \nabla \log p(y)^\top \boxed{\nabla_x k(x,y)} \\ &+ \nabla \log p(x)^\top \boxed{\nabla_y k(x,y)} + \langle \boxed{\nabla_x k(x,\cdot)}, \boxed{\nabla_y k(\cdot,y)} \rangle_{\mathcal{F}^d} \end{aligned}$$

Methods: Main Results

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z')$$

Stuff with **kernels** and its gradients

Methods: Main Results

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z')$$

Stuff with **kernels** and its gradients

Theorem: Let *p*, *q* be probability measure, *Z* ~ *q*, then under certain conditions (finite math. expectations...):

$$S_p(Z) = 0 \quad \Longleftrightarrow \quad p = q$$

Methods: Main Results

$$S_p(Z) := \sup_{\|f\|<1} \mathbb{E}\,(T_p f)(Z) - \mathbb{E}\,(T_p f)(X)$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z')$$

Theorem: Let *p*, *q* be probability measure, *Z* ~ *q*, then under certain conditions (finite math. expectations…):

$$\boxed{S_p(Z) = 0 \quad \Longleftrightarrow \quad p = q}$$

Stain discrepancy – indicator of similarity!

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z')$$

was shown 2 slides ago

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z') \longrightarrow nV_n = \frac{1}{n} \sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

was shown 2 slides ago

estimator

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z') \longrightarrow nV_n = \frac{1}{n} \sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

But what if Z_i exhibit **correlation** behaviour?

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E}h_p(Z, Z') \longrightarrow nV_n = \frac{1}{n}\sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

But what if Z_i exhibit correlation behaviour? **The Wild Bootstrap technique**

Markov chain: $W_{t,n} = \mathbf{1}(U_t > a_n)W_{t-1,n} - \mathbf{1}(U_t < a_n)W_{t-1,n}$

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E} h_p(Z, Z') \longrightarrow nV_n = \frac{1}{n} \sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

But what if Z_i exhibit correlation behaviour? **The Wild Bootstrap technique**

Markov chain: $W_{t,n} = \mathbf{1}(U_t > a_n)\boxed{W_{t-1,n}} - \mathbf{1}(U_t < a_n)\boxed{W_{t-1,n}}$

Methods: Bootstrap

$$H_0 : S_p(Z) = 0 \quad \text{vs} \quad H_1 : S_p(Z) \neq 0$$

$$S_p(Z)^2 = \mathbb{E}h_p(Z, Z') \longrightarrow \boxed{nV_n = \frac{1}{n}\sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)}$$

But what if Z_i exhibit correlation behaviour? **The Wild Bootstrap technique**

Markov chain: $W_{t,n} = \mathbf{1}(U_t > a_n)W_{t-1,n} - \mathbf{1}(U_t < a_n)W_{t-1,n}$

$$\boxed{nB_n = \frac{1}{n}\sum_{i,j=1}^{n} W_{i,n}W_{j,n}h\left(Z_i, Z_j\right)}$$

Experiments
Experiments
Experiments

Experiments

Most interesting part, u know:)

Experiments
Experiments
Experiments
Experiments

Experiment #1

Student’s t-distribution **vs** Normal

$$H_0 : Z \sim \mathcal{N}(0,1) \quad \text{vs} \quad H_1 : Z \not\sim \mathcal{N}(0,1)$$

Experiment #1

Student’s t-distribution **vs** Normal

$$H_0 : Z \sim \mathcal{N}(0,1) \quad \text{vs} \quad H_1 : Z \not\sim \mathcal{N}(0,1)$$

$$nB_n = \frac{1}{n} \sum_{i,j=1}^{n} W_{i,n} W_{j,n} h\left(Z_i, Z_j\right)$$

Markov chain: $W_{t,n} = \mathbf{1}(U_t > a_n) W_{t-1,n} - \mathbf{1}(U_t < a_n) W_{t-1,n}$

Experiment #1

Student's t-distribution **vs** Normal

$$H_0 : Z \sim \mathcal{N}(0,1) \quad \text{vs} \quad H_1 : Z \not\sim \mathcal{N}(0,1)$$

$$nB_n = \frac{1}{n}\sum_{i,j=1}^{n} W_{i,n}W_{j,n}h\left(Z_i, Z_j\right)$$

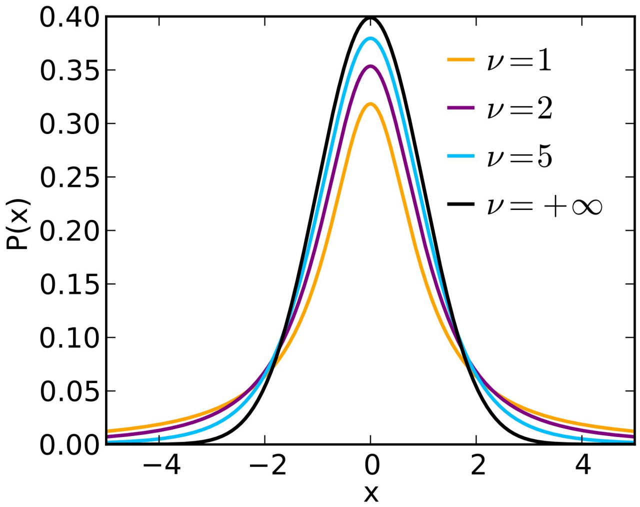
Markov chain: $W_{t,n} = \mathbf{1}(U_t > \boxed{a_n})W_{t-1,n} - \mathbf{1}(U_t < a_n)W_{t-1,n}$

How to choose?

Experiment #1

Student's t-distribution **vs** Normal

$$H_0 : Z \sim \mathcal{N}(0,1) \quad \text{vs} \quad H_1 : Z \cancel{\sim} \mathcal{N}(0,1)$$

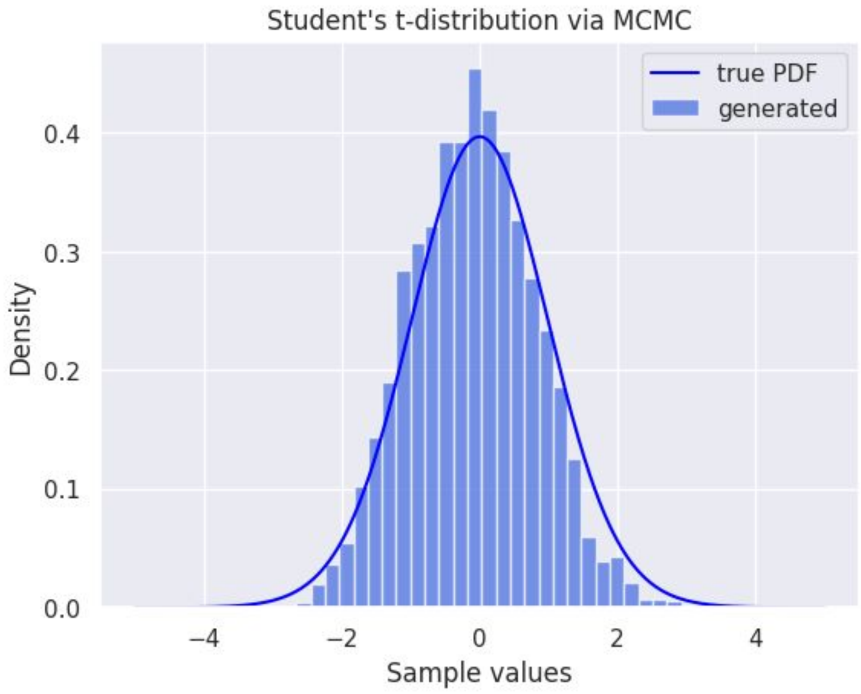


1. Make a sample from **Student's t-distribution** (going to Normal distribution with $\nu \to \infty$)

2. Expect **low-p-values** when degrees of freedom are small

Experiment #1

Student’s t-distribution **vs** Normal



- Sampled using Markov Chain Monte Carlo

- Distribution **PDF**:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Experiment #1

Student’s t-distribution **vs** Normal

```
for dof in degrees_of_freedom:
    for n in range(N_exp):
        X = t_student_distrib(5000, dof, 0.01)
        test = GaussianQuadraticTest(grad_log_normal)
        V_n, _ = test.get_statistics(X)
        p_value = test.compute_pvalues(V_n)
```

generate t-student

Experiment #1

Student's t-distribution **vs** Normal

```
for dof in degrees_of_freedom:
    for n in range(N_exp):
        X = t_student_distrib(5000, dof, 0.01)
        test = GaussianQuadraticTest(grad_log_normal)
        V_n, _ = test.get_statistics(X)
        p_value = test.compute_pvalues(V_n)
```

compute V-statistics

$$nV_n = \frac{1}{n}\sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

Experiment #1

Student's t-distribution **vs** Normal

```
for dof in degrees_of_freedom:
    for n in range(N_exp):
        X = t_student_distrib(5000, dof, 0.01)
        test = GaussianQuadraticTest(grad_log_normal)
        V_n, _ = test.get_statistics(X)
        p_value = test.compute_pvalues(V_n)
```

compute V-statistics

$$nV_n = \frac{1}{n} \sum_{i,j=1}^{n} h\left(Z_i, Z_j\right)$$

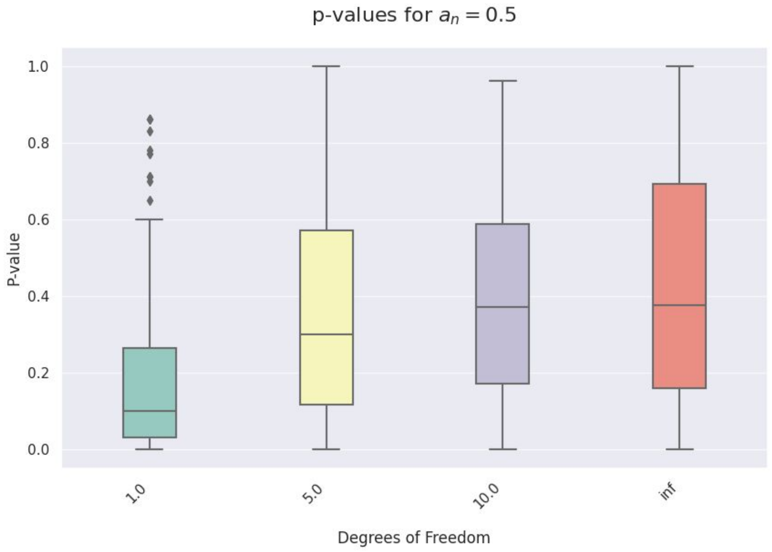
compute p-values

$$nB_n = \frac{1}{n} \sum_{i,j=1}^{n} W_{i,n} W_{j,n} h\left(Z_i, Z_j\right)$$

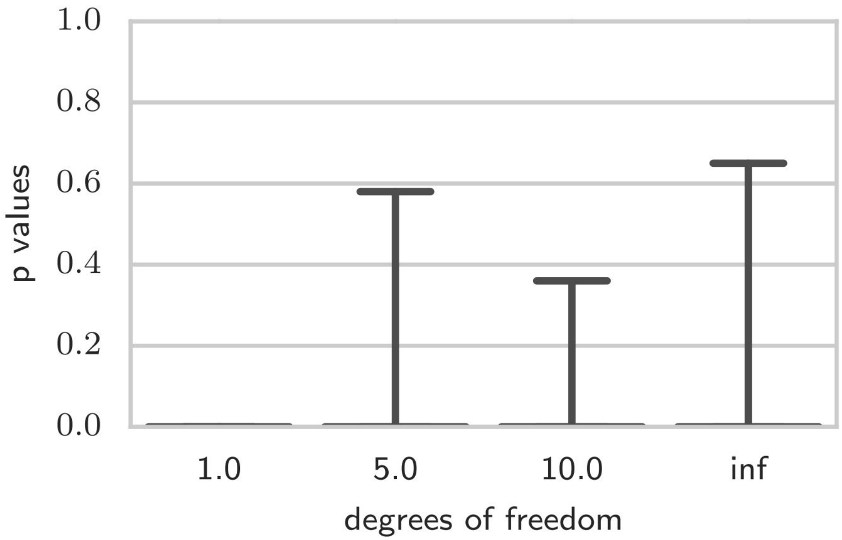
```
count(nBn > nVn)
```

Experiments: $a_n = 0.5$

Graph we **obtained**

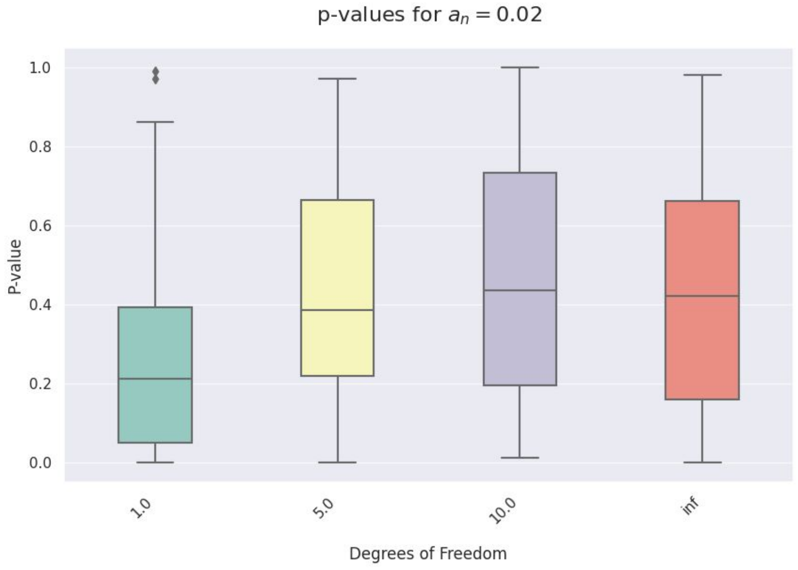


Graph from the **article**

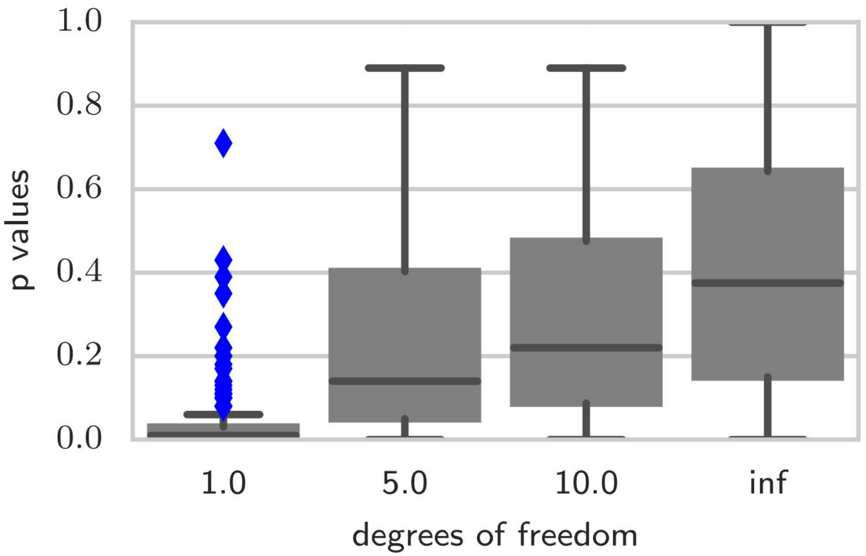


Experiments: $a_n = 0.02$

Graph we **obtained**

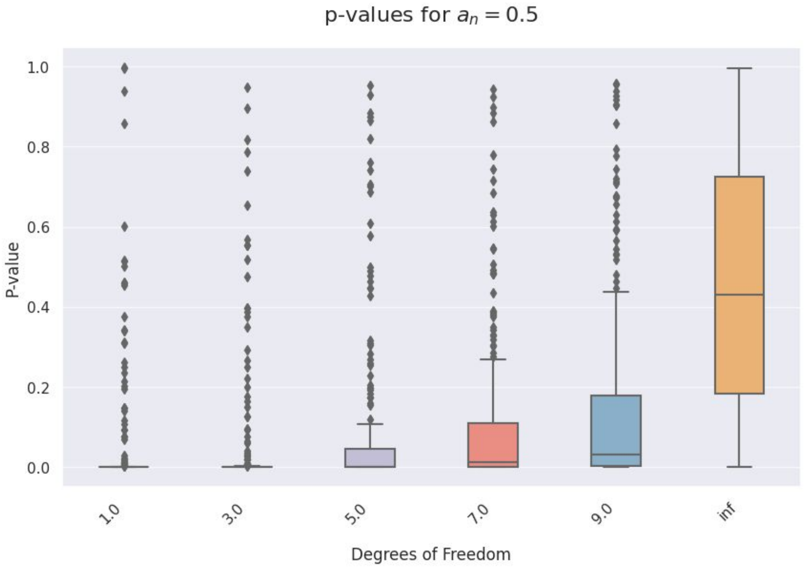


Graph from the **article**

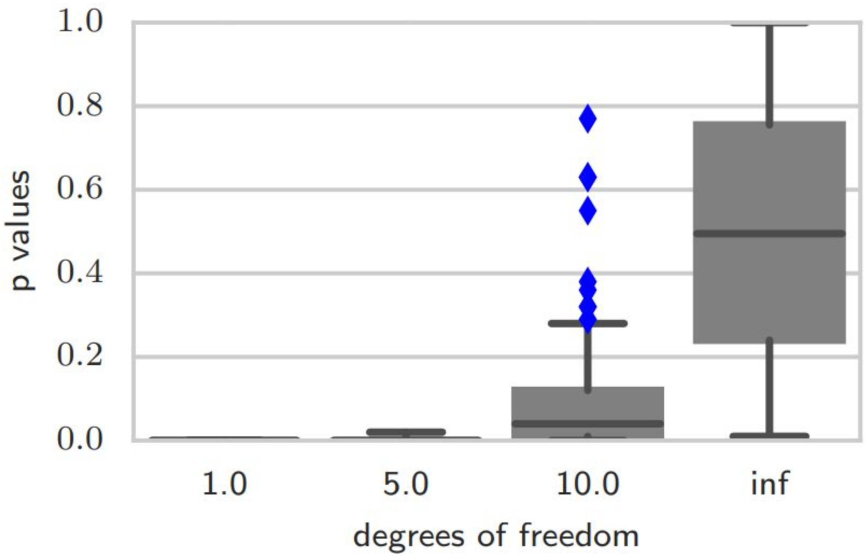


Experiments: thinning

Graph we **obtained**

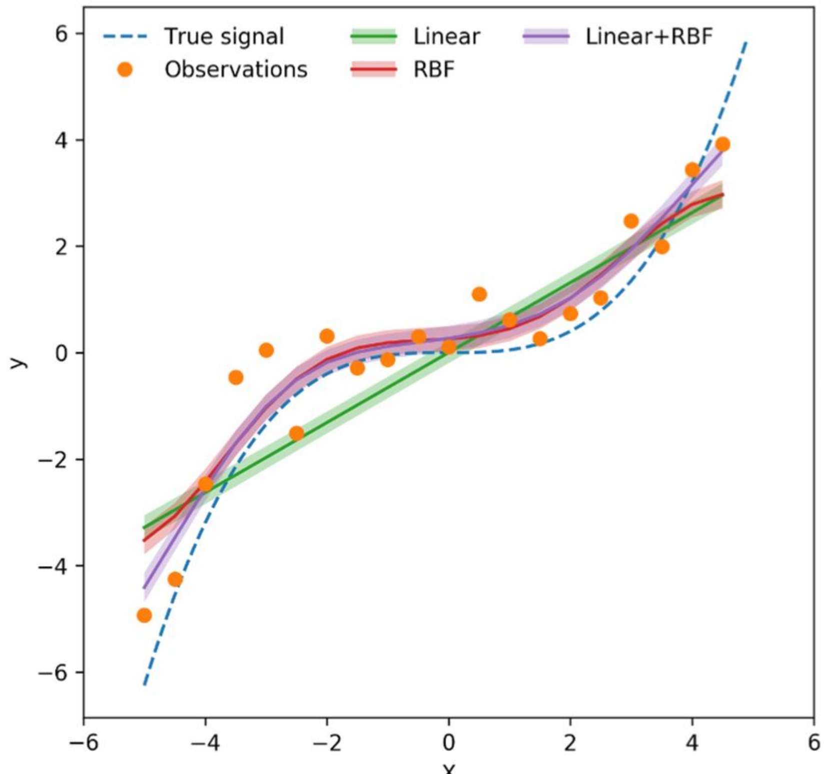


Graph from the **article**



Experiment #2

Statistical model criticism on gaussian processes



Kernel selection.

Predictions made by GPR when using the Linear, RBF kernels.

The shaded region around each curve represents the 95% CI

Experiment #2

Statistical model criticism on gaussian processes

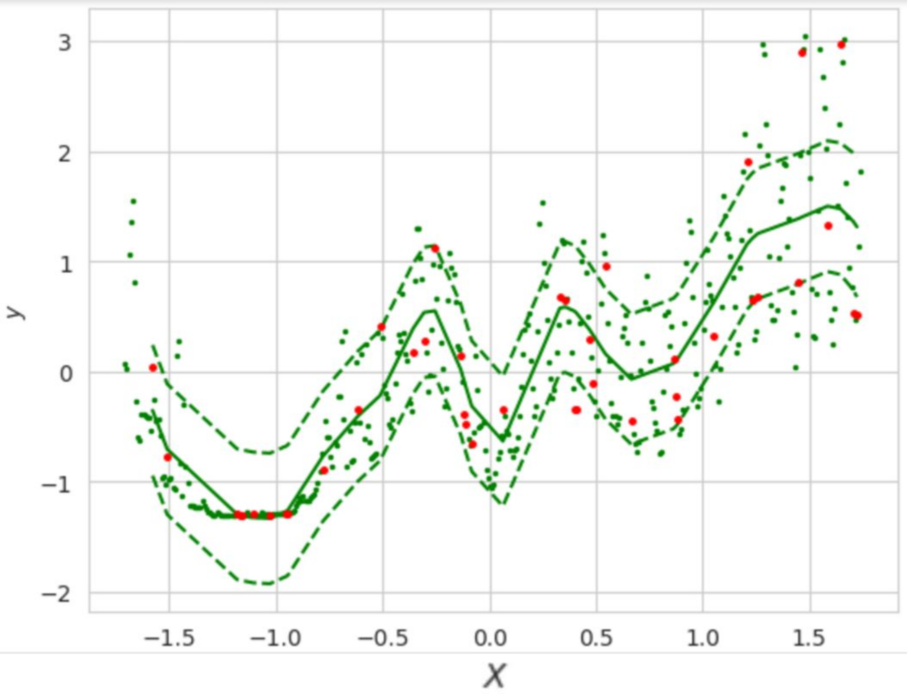
- **Solar** dataset
- 1D regression problem with N=402
- We fit N_{train} = 361 data using a GP with a **squared exponential kernel** and a Gaussian noise model

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right)$$

- H_0 : solar dataset ~ predictive distribution

Experiment #2

Statistical model criticism on gaussian processes

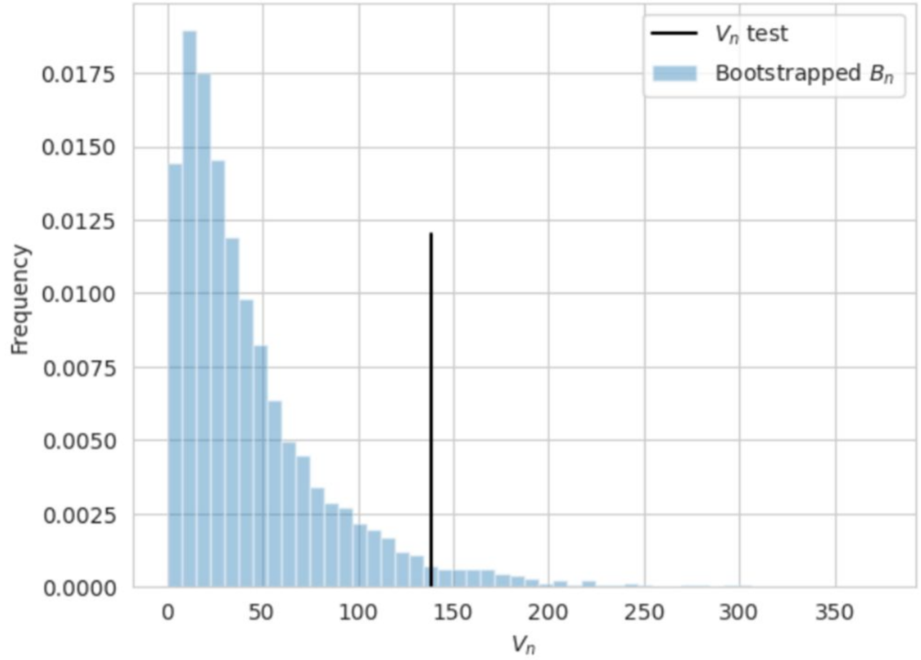


Fitted GPR:

- **Green dots** are train dataset
- **Red dots** are test dataset
- Green line is GPR predicted line
- Dotted green lines are left and right edges of confidence interval

Experiment #2

Statistical model criticism on gaussian processes



1. Bootstrapped B_n distribution with the test statistic V_n marked.
2. That it is **unlikely** that the test points were generated by the fitted GP model.

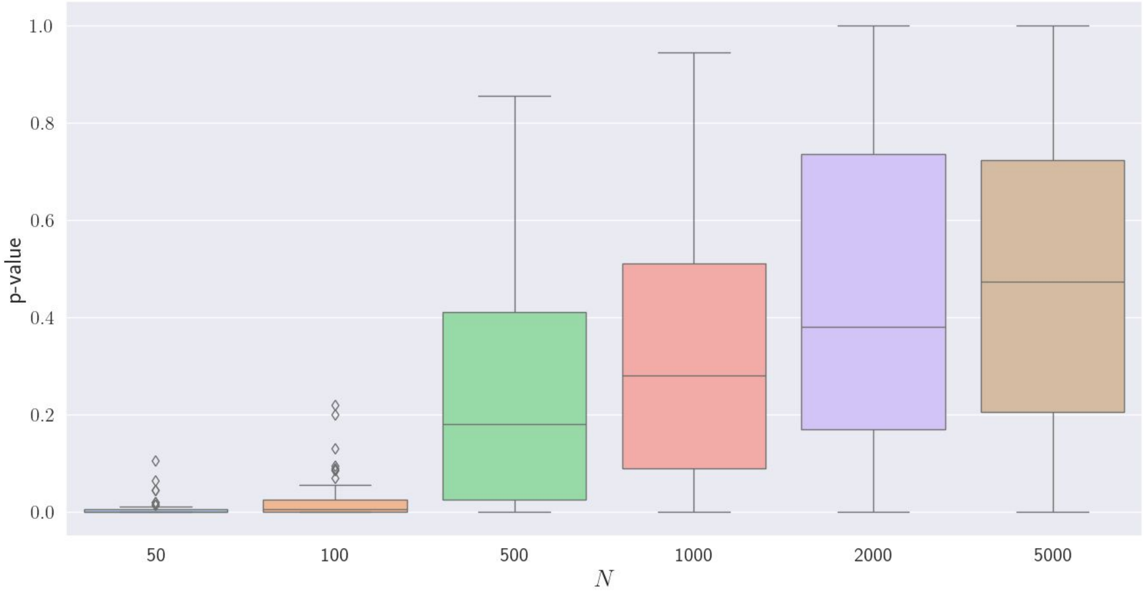
Experiment #3

Convergence in non-parametric density estimation

- Measuring quality-of-fit nonparametric density estimation
- 2 density models:
  - The infinite dimensional exponential family
  - The approximation to this model via random Fourier features

Experiment #3.1

Convergence in non-parametric density estimation

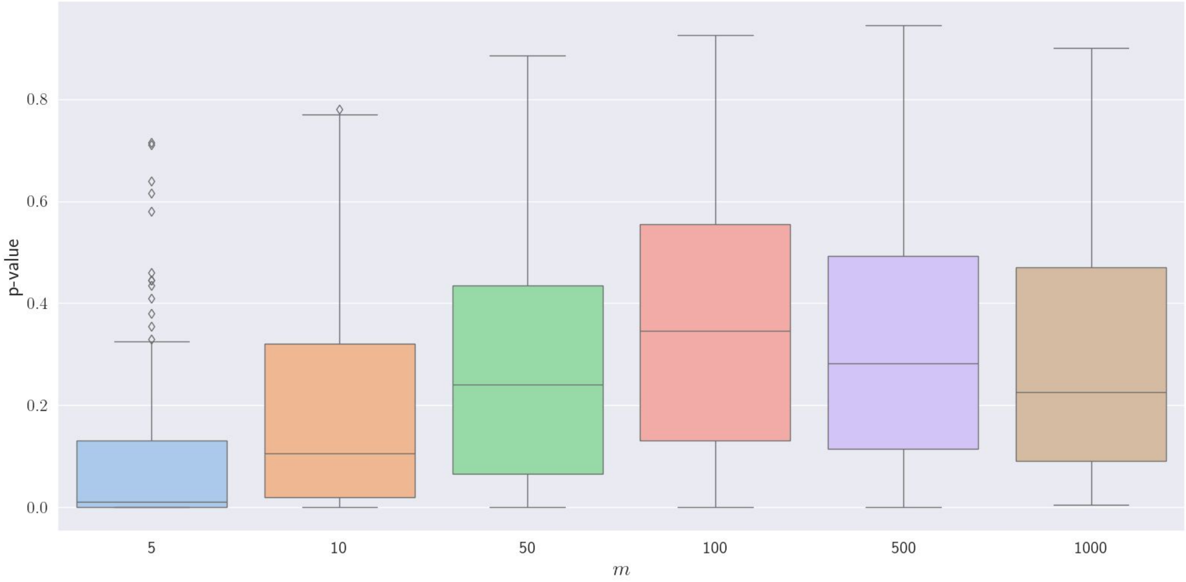


Distribution of p-values

- N observations
- A quadratic time test on N_{test} = 500
- Goal: identify N sufficiently large, that the method **would not** reject the null hypothesis

Experiment #3.2

Convergence in non-parametric density estimation



Distribution of p-values

- F is approximated by a finite dictionary of **random Fourier features**
- The same N number is used
- P-values **do not** have a uniform distribution, even for a large number of random features

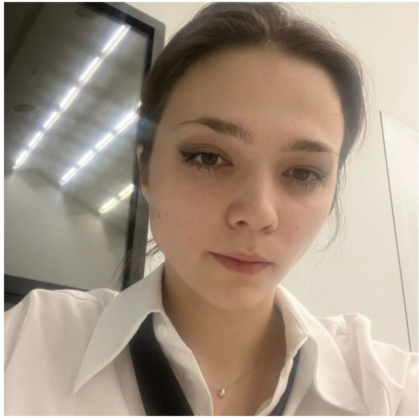
Conclusion
Conclusion
Conclusion

Conclusion

Let’s recap what we have done

Conclusion
Conclusion
Conclusion
Conclusion

Contribution of Team members

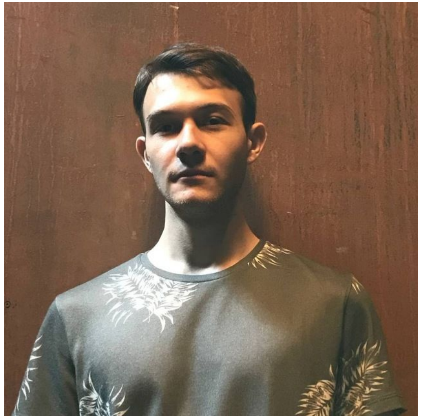


Viktoriia Zinkovich
Data Science, MS-1

- **Experiment:** Student's t-distribution VS normal
- Presentation design

Kamil Garifullin
Data Science, MS-1

- **Experiment**: Statistical Model criticism on Gaussian Processes
- Presentation design



Maksim Osipenko
Data Science, MS-1

- **Experiment**: Convergence in non-parametric density estimation
- Problem statement

Conclusion

- Construction of the RKHS-based Stein discrepancy and associated statistical test

- Experimental illustrations on synthetic examples:

  student’s t vs normal

  statistical model criticism

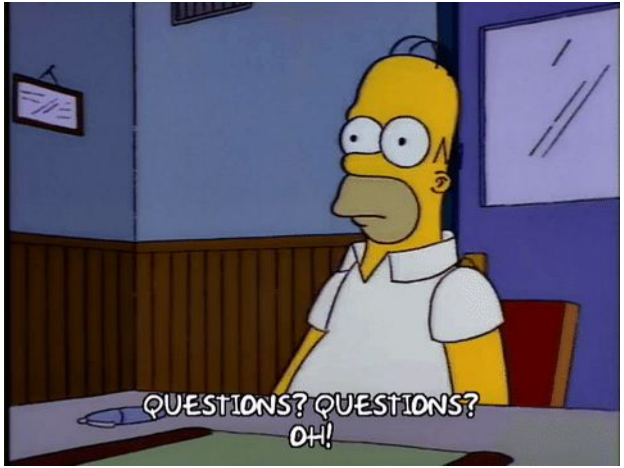
  convergence in nonparametric density estimation.

Questions?

Maksim Osipenko
Maksim.Osipenko@skoltech.ru
Data Science

Viktoriia Zinkovich
Viktoriia.Zinkovich@skoltech.ru
Data Science

Kamil Garifullin
Kamil.Garifullin@skoltech.ru
Data Science



Code is available at Github

[1] **A Kernel Test of Goodness of Fit**. Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton