Introduction to Nonparametric Statistics

Paweł Polak

January 26, 2016

STAT W4413: Nonparametric Statistics - Lecture 3

What is nonparametric statistics about?

We will discuss a few examples of parametric and nonparametric estimation/testing and compare their advantages and disadvantages.

Example: Test under Gaussian assumption

Suppose that $X_1, X_2 \ldots, X_n \overset{i.i.d.}{\sim} N(\zeta, 1)$.

$$H_0 : \zeta = \zeta_0 \qquad \text{vs.} \qquad H_1 : \zeta \neq \zeta_0. \tag{1}$$

One way to test H_0 is to compare the sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ with ζ_0.

We expect that if the null hypothesis is true then $\bar{X}$ should be close to ζ_0. Therefore, we suggest the following test:

$$\begin{aligned} &\text{Reject } H_0 && \text{if } |\bar{X} - \zeta_0| > \kappa, \\ &\text{Accept } H_0 && \text{if } |\bar{X} - \zeta_0| \leq \kappa, \end{aligned} \tag{2}$$

We need to set the parameter κ, that is called the threshold parameter here. This parameter affects the performance of the test.

Performance criteria for tests

There are two important criteria that measure the performance of the test. To understand these two criteria check the following table:

		Truth	
		H_0	H_1
Test decision	H_0	Good decision	Type II error
	H_1	Type I error	Good decision

- *Type I error* occurs when the null hypothesis is correct, but our test rejects it. **When $\kappa \ll 1$**
- *Type II error* occurs when the alternative is true, but we accept the null. **When $\kappa \gg 0$**
- These two errors are usually in trade-off, meaning that when the probability of one of them increases, the probability of the other one decreases, and vice versa.
- In the test on the previous slide, as we increase the value of κ we accept H_0 on a larger region, and reject it on a smaller region.
- Therefore, as κ increases, the probability of Type I error decreases, while the probability of Type II error increases. Hence, these two probabilities are considered as two measures of performance for the tests.
- The probability of Type I error is also known as the *significance level*.

Setting test parameters

- As mentioned in the last slide the parameter κ changes both the significance level and probability of Type II error.
- In many cases, we prefer to be on the conservative side and do not reject the null hypothesis unless we have strong evidence against it.
- In such cases, we set the probability of type I error or significance level to a small number α (e.g. 0.01), and set the parameter κ accordingly.

Let's follow this strategy here and calculate the parameter κ....

Setting test parameters

Suppose that H_0 is correct. Then $\bar{X} \sim N(\zeta_0, \frac{1}{n})$. Therefore our test statistic $\bar{X} - \zeta_0 \sim N(0, \frac{1}{n})$, and the probability of Type I error is equal to

$$\mathbb{P}(|\bar{X} - \zeta_0| > \kappa | H_0) = 2\mathbb{P}(\bar{X} - \zeta_0 < -\kappa | H_0) = 2\Phi(-\kappa\sqrt{n}),$$

where $\Phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ is the cdf of a standard Gaussian random variable. If we set the significance level to α, we obtain $\kappa = -\frac{\Phi^{-1}(\alpha/2)}{\sqrt{n}}$

Remark Once we set κ it is straightforward to calculate the probability of Type II error for any given value of $\zeta \neq \zeta_0$.

Parametric versus Nonparametric

- Our discussion so far falls in the category of parametric statistics.
- In fact we start with a parametric model for our data, e.g., the data follows a Gaussian distribution with certain mean and our test checks some conditions on the parameter or parameters of those distributions.
- Nonparametric statistics course challenges this assumption. The main questions we would like to answer in this course are the following:
  - What if we do not know the actual distribution?
  - How should we design tests and when shall we use them?

In the next slides we review a few examples of nonparametric tests.

Goodness of fit

- Consider again the Gaussian example from the last slides.
- The first question we can ask about the problem set-up is whether the Gaussian distribution is the right distribution for the data or not.
- This is the type of question that is known as "goodness of the fit".
- Here is a more formal statement of the problem:
  Suppose that the data $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F(x)$. The only assumption we have on F is that it is continuous. We believe that F is a certain distribution F_0 ($N(0,1)$ for example). Therefore we would like to test $H_0 : F(x) = F_0(x)$ for every x against the alternative $H_1 : F(x) \neq F_0(x)$ for at least one value of x. Note that this test does not make any parametric model on F and hence it falls in the category of non-parametric tests.
- The following simple lemma will let us design a famous Kolmogorov-Smirnov test:

Goodness of fit

Lemma

Let $X \sim F$. Define the indicator function as

$$\mathbb{I}(x \leq a) \triangleq \begin{cases} 1 & x \leq a, \\ 0 & x > a. \end{cases} \tag{3}$$

Then

$$F(a) = \mathbb{E}(\mathbb{I}(X \leq a)).$$

Proof.

Let f be the pdf of X. Then by definition

$$\mathbb{E}(\mathbb{I}(X \leq a)) = \int_{-\infty}^{\infty} \mathbb{I}(x \leq a) f(x) dx = \int_{-\infty}^{a} f(x) dx = F(a).$$

□

Goodness of fit

- This lemma shows that $F(a)$ can be expressed as the expectation of the indicator function.

- Therefore, in order to estimate $F(a)$, one simple way is to replace the expectation with the sample mean, i.e.,

$$\hat{F}(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq a). \tag{4}$$

- We can estimate $F(a)$ for every value of a and obtain the *empirical distribution function.*

Example

Example

Let the observed samples of a certain distribution be $0.9, 1.65, 0.49, -1, 0.6, -0.95, -0.25, 2.1, -0.9, 0.53$. The empirical distribution of this data is shown in the next slide. Once we have an estimate of the CDF, we can easily design tests to check the validity $F(x) = F_0(x)$. If $\hat{F}(a)$ is "close" to F_0 then the null should be accepted and if $\hat{F}$ is "far from" F_0, then H_0 should be rejected. The only statement that has been remained vague, is the measure of "closeness" or "farness". Different measures may lead to different tests and we will go over them in detail. Here, we briefly mention one measure that leads us to Kolmogrov-Smirnov test. *Kolmogorov-Smirnov* test is based on the following criteria:

$$\text{Reject} \quad H_0 \qquad \text{if} \qquad D^* := \sup_{a \in \mathbb{R}} |\hat{F}(a) - F_0(a)| > \kappa.$$

We will measure the performance of this test and describe its limitations later in the course.

D* is Distribution Free

Goodness of fit

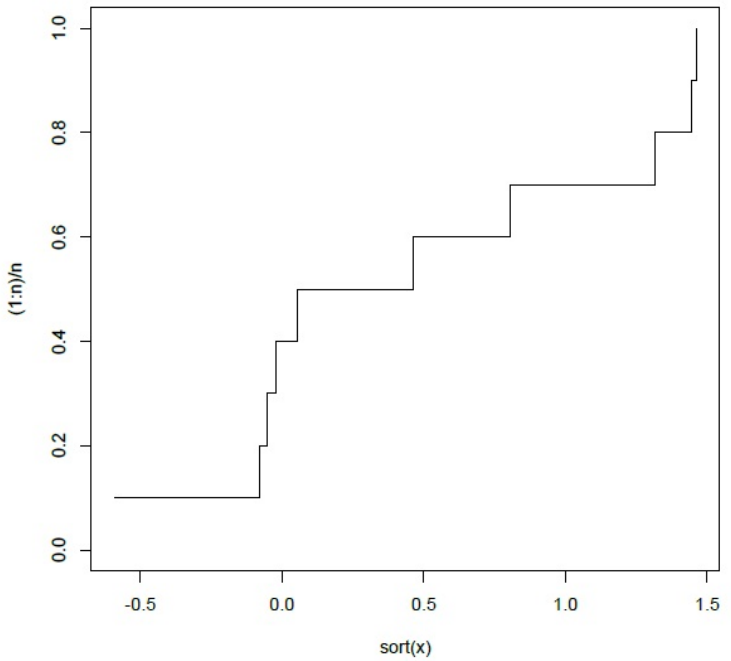

Figure : The empirical cumulative distribution function of the data in Example above

Sign test

- Now suppose that we have performed "goodness of the fit test" on the data and are convinced that the Gaussian model we chose for it is not correct.
- But still we are interested to test the mean of the data. How can we do this test?
- The first idea that comes to mind is to use the same test that we had in (2).
- Is there going to be any problem? Before we proceed further let us formally state this problem.

Example

Let $X_1 \ldots, X_n \overset{iid}{\sim} F(x)$, where F is not known. Define $\mu \triangleq \mathbb{E}(X_i)$. We would like to test for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Sign test

- Let's consider the test in (2).

  `if dist symmetric, use median`

- The first problem arises when we want to calculate the value of κ.
- Since the probability distribution is not known, we cannot calculate the probability of Type I error and therefore, we cannot set κ properly.
- But, we can prove that the problem here is more severe than only calculating the probability of Type I or Type II errors.
- In fact, there is no "efficient test" (More formally this means that for every test, if the significance level is α, then the probability of Type II error is larger than $1 - \alpha$.) [1]
- The main message of Example 2 is that if we do not know the distribution, we will have some limitation in terms of the questions we can address.
- Therefore, it is important to learn the types of questions we can answer, so that we can translate our problems into these forms.
- For instance, if in Example 2 we replace the mean by the median, then we can design tests for that.

[1] Check Chapter 15 of "Testing statistical hypotheses" by Lehman and Romano for further discussion. for testing the value of the mean under the settings of example 2. This is beyond the scope of this course.

Sign test

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x)$ where F is a continuous CDF. Let $\mu = MED(X_i) < \infty$.
- We would like to check if μ is equal to μ_0 or not.
- However, we state the problem in slightly different way.
- Define $p \triangleq \mathbb{P}(X \leq \mu_0)$.
- If μ_0 is the median then we expect p to be equal to $1/2$, otherwise $p \neq 1/2$.
- This leads us to the following hypothesis check:

$$H_0 : p = \frac{1}{2} \qquad \text{versus} \qquad H_1 : p \neq \frac{1}{2}.$$

- Designing test for this hypothesis is very similar to designing tests for the goodness of the fit. With a similar approach we can estimate p from the data. Define this estimate as

$$\hat{p} \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq \mu_0).$$

- Clearly, if μ_0 is the median of the data, we would expect $\hat{p}$ to be close to 0:5. Therefore, our test will be

$$\left| \hat{p} - \frac{1}{2} \right| \geq \gamma \quad \text{then reject} \quad H_0.$$

- The main question that we have not answered yet is that whether under the general settings we are interested here, i.e., arbitrary distribution F, we can measure the probability of Type I and Type II errors?

Sign test

- Let's start with the significance level.
- First note that $\mathbb{I}(X_i \leq \mu_0)$ is a discrete random variable that can only take on two different values $\{0, 1\}$.
- Under the null hypothesis we have

$$\mathbb{I}(X_i \leq \mu_0) = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ 0 & \text{with probability } \frac{1}{2}. \end{cases} \tag{5}$$

- Therefore, under H_0

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}(X_i \leq \mu_0) = k\right) = \binom{n}{k} \left(\frac{1}{2}\right)^n.$$

Sign test

- Now we can calculate the probability of Type I error.

$$\begin{aligned}
\mathbb{P}\left(\left|\hat{p}-\frac{1}{2}\right| \geq \gamma \mid H_0\right) &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}(X_i \leq \mu_0) \geq n(\gamma+\frac{1}{2})\right) \\
&+ \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{I}(X_i \leq \mu_0) \leq n(\frac{1}{2}-\gamma)\right) \\
&= \sum_{k=\lceil n(\gamma+1/2)\rceil}^{n} \binom{n}{k}\left(\frac{1}{2}\right)^n + \sum_{k=0}^{\lfloor n(\frac{1}{2}-\gamma)\rfloor} \binom{n}{k}\left(\frac{1}{2}\right)^n,
\end{aligned}$$

  where for any $a \in \mathbb{R}$ $\lfloor a \rfloor$, $\lceil a \rceil$ denote the largest integer that is less than or equal to a and the smallest integer that is larger than or equal to a, respectively. Similarly, we can calculate the probability of Type II error for any value of μ.

As you can see here by switching the problem from mean to median and presenting it in a slightly different way we have been able to design a test.

Parametric versus nonparametric tests

- The examples we have seen so far show the strengths and weaknesses of nonparametric methods.
- The main advantage of these methods is that they don't impose any structure on the data.
- So, in cases we don't have any information about the data, nonparametric methods will be useful. Specially, in cases where we would like to impose certain structure on the data (such as Gaussianity) to simplify the problem/solution, nonparametric methods enable us to evaluate the accuracy of the model we have assumed.
- For instance Kolmogorov-Smirnov test checks whether the distribution is Gaussian.
- The benefits of the nonparametric methods come at a price:
    - First, as we mentioned in Example 2, some of the hypotheses we would like to test, do not have "efficient" tests in nonparametric settings. So, we should be careful in designing tests on the data.
    - Second, as we will see later the generality of nonparametric tests reduces their power in some cases. We will get into this issue later.
- We will explain in detail when/where you should use nonparametric or parametric tests in this course.

Curve fitting

Nonparametric statistics includes another set of problems that are different in nature from nonparametric testing. The so called *curve fitting* problems.

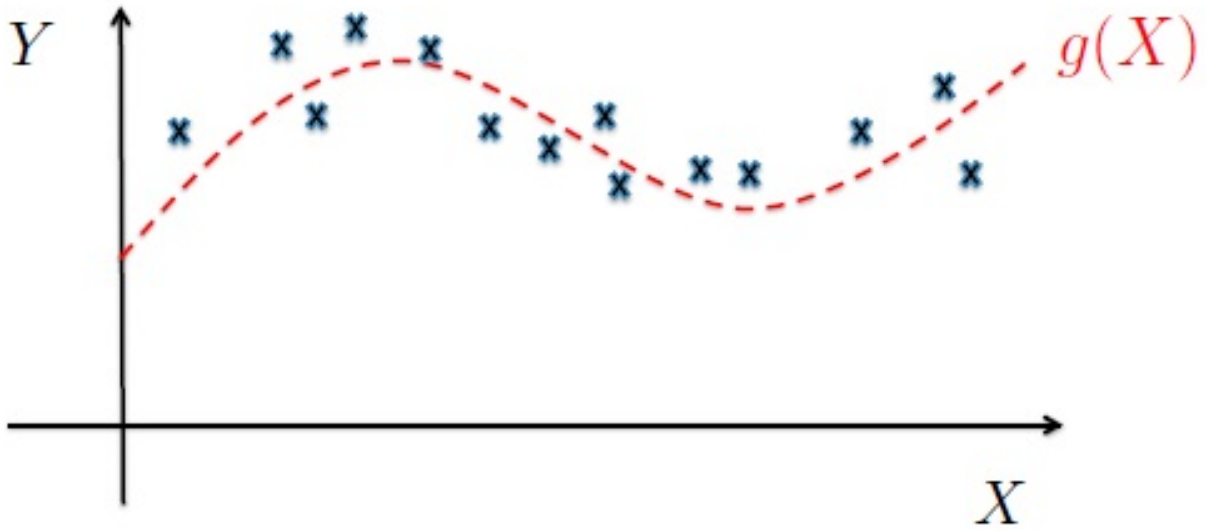

Figure : Curve fitting problem. The crosses are data points. The goal is to estimate the red curve $g(X)$.

Maximum likelihood principle

We have observed $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. We believe that Y samples are a function of the X samples, i.e.,

$$Y_i = g(X_i) + Z_i,$$

where Z_i denotes the noise. You may have different models for the noise. But for this discussion assume that the noise is Gaussian. Our goal is to characterize or estimate the function g. Figure 2 illustrates this problem.

The strategy that we have learned in "linear regression" or inference courses follows these steps:

1. Consider a model for g. Let's first assume that $g = \alpha_0 + \alpha_1 X$, where only α_0 and α_1 are unknown. This is also called parametric model, as we have converted the problem into the problem of estimating a finite number of parameters.

2. Let the noise be Gaussian $N(0, \sigma^2)$.

3. Use maximum likelihood principle to estimate α_0 and α_1. If the estimates are $\widehat{\alpha}_0, \widehat{\alpha}_1$, then our final estimate of g would be $\hat{g} = \widehat{\alpha}_0 + \widehat{\alpha}_1 X$.

Maximum likelihood principle

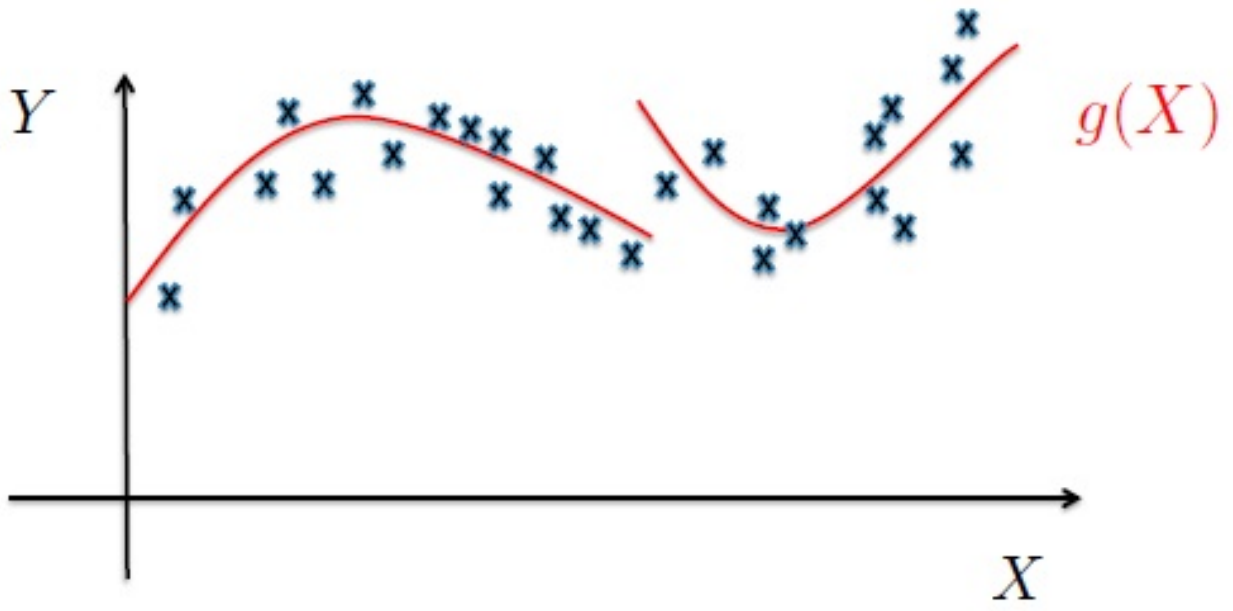

Figure : Curve fitting problem. The crosses are data points. The goal is to estimate the red curve $g(X)$.

While this approach is very powerful and works well in some applications, in many other applications g cannot be modeled as a linear function.

What can we do in those situations?

Maximum likelihood principle

- The parametric approach is to come up with another model that can describe more complicated functions.
- One example would be to model g with a polynomial with a higher degree. For instance, $g(X) = \sum_{i=0}^{p-1} \alpha_i X^i$.
- Then we use the same strategy to calculate an estimate of $\alpha_0, \ldots, \alpha_{p-1}$.
- Again this strategy is called parametric, since we have converted the problem of estimating g into the problem of estimating finite number of parameters.

Nonparametric curve fitting

- In many cases polynomials are not rich enough to describe the functions.
- For instance, we may consider the case where the function g is not continuous, or it is continuous but not differentiable.
- In all these cases the polynomial model does not work.
- Nonparametric methods consider general forms of functions and impose minimal restriction on the functions.
- Then they come up with new approaches to estimate the function g.
- For instance, suppose we assume that the function is piecewise differentiable.
- Note that we do not even impose constraint on the location of "non-differentiablity" points and/or the number of such points.
- This class of functions seems to be rich enough to cover the functions we see in many applications.

But can we estimate such functions?

Nonparametric curve fitting

Here is a very simple idea:

- Suppose that $\max(X_1, X_2, \ldots, X_n) = 1$ and $\min(X_1, X_2, \ldots, X_n) = 0$
- Then break the $[0, 1]$ interval into smaller intervals of size δ as depicted in Figure 4.
- Since δ is small, in the smooth regions the function "looks" linear.
- Therefore, we can estimate the function in each interval with a line, independent of the other parts.
- This gives a reasonably good estimate. But, as you can imagine it is very naive.
- We will see more advanced methods and will analyze their performance both theoretically and empirically later in the course.