Nonparametric statistics: a teaser

(notes by Pavel Chigansky)

In statistics the data $X \in \mathcal{X}$ is assumed to be sampled from an unknown probability measure $\mathbb{P}$. The objective is to infer about $\mathbb{P}$ given the observed sample X. To this end a collection of probabilities $\mathcal{P}$, is postulated and the true unknown probability measure $\mathbb{P}$ is assumed to belong to $\mathcal{P}$. This collection is called the *statistical model*. Inference about $\mathbb{P}$ is done by fitting the data to the probability measures from the model.

In parametric models, the collection $\mathcal{P} = (\mathbb{P}_\theta)_{\theta\in\Theta}$ is indexed by a *parameter* variable θ, which takes values in a finite dimensional *parameter space* Θ, such as $\Theta = \mathbb{R}^d$ for some $d \in \mathbb{N}$. In nonparametric models, $\mathcal{P}$ is still parameterised by $\theta \in \Theta$, but the parameter space Θ is *infinite dimensional*, typically a space of relevant functions or sequences. The goal is to design methods of statistical inference with provable performance guarantees. Here are some distinctions between parametric and nonparametric models.

(1) In parametric statistics there are generic methods, which perform optimally at least in the large sample limit. Common examples are the likelihood based methods, such as the Maximum Likelihood and the Bayes estimators, or the Generalized Likelihood Ratio Test. These procedures do not apply directly to infinite dimensional parameter spaces, and in fact, much of the nonparametric methodology is based on completely different ideas.

(2) How many samples are needed to achieve a desired accuracy? In parametric estimation problems, the error cannot decrease faster than $n^{-1/2}$, if the model is sufficiently regular, and this optimal rate is attained by the ML and the Bayes estimators with positive prior densities. This is quite robust and, in particular, the optimal rate of $n^{-1/2}$ is not affected by the dimension of Θ (as long as it is finite!), the particular loss function which defines the risk, etc. In nonparametric models the picture is completely different. Typically slower rates are possible, depending on a priori properties of the unknown functions, such as smoothness, monotonicity, etc. and the risk functional.

lecture notes for "Advanced Statistical Models B" course.

Below some common nonparametric models are surveyed in the context of classical problems of statistical inference: point and set estimation, and hypothesis testing.

1. Point estimation

Point estimation is concerned with finding a statistic $\widehat{\theta} : \mathcal{X} \mapsto \Theta$, called the *estimator*, whose value at the available sample is interpreted as an estimate (a guess) of the true unknown value of the parameter or some known function of it.

1.1. Distribution and functionals estimation. Suppose we observe an i.i.d. sample $X = (X_1, ..., X_n)$ from an unknown cumulative distribution function (c.d.f) F on $\mathbb{R}$, which is to be estimated from X. Let $\mathbb{P}_F$ be the product probability measure on $\mathcal{X} = \mathbb{R}^n$, whose one-dimensional marginals have distribution function F, so that $X \sim \mathbb{P}_F$. Keep in mind that both X and $\mathbb{P}_F$ change with n, this dependence will be omitted for brevity from the notations.

An estimator for F is a statistic $\widehat{F}_n$, which qualifies as a legitimate distribution function. A reasonable estimator is the *empirical distribution function*

זה בעצם היסטוגרמה באופן פורמלי

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{X_j \le x\}}, \quad x \in \mathbb{R}.$$

Indeed this is a nondecreasing piecewise constant function, which equals 0 and 1 for all sufficiently small and large x's respectively. It is a purely discrete distribution function with atoms at X_j's, even if the true unknown distribution is continuous.

Often in applications, it is required to estimate some *functional* of F, rather than F itself. A functional is a map from the space of distributions to some finite dimensional space, such as $\mathbb{R}^k$. A few common examples are

$$\begin{aligned} \mu_F &= \int_{\mathbb{R}} x dF(x), && \text{(mean)}, \\ \sigma_F^2 &= \int_{\mathbb{R}} (x-\mu_F)^2 dF(x), && \text{(variance)}, \\ q_F(p) &= \inf\Big\{x \in \mathbb{R} : F(x) \ge p\Big\}, && (p\text{-quantile}), \end{aligned} \tag{1.1}$$

where the integrals are understood in the Lebesgue-Stiltjes sense. If the functional of interest is well defined on purely discrete distributions, a reasonable estimator can be obtained by plugging $\widehat{F}_n$ in place of F. For example, such plug-in or *substitution* estimator for the mean is the familiar empirical mean statistic,

$$\mu_{\widehat{F}_n} = \int_{\mathbb{R}} x d\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}_n.$$

1.2. Density estimation. In the setting as above, i.e. $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} F$ we may be interested in estimating the probability density

$$f(x) = \frac{d}{dx}F(x),$$

ההיסטוגרמה היא לא גזירה
לכן יש צורך באומדנים אחרים

assuming that it exists. A density estimator $\widehat{f_n}$ is a nonnegative function, which integrates to 1 over $\mathbb{R}$, constructed using the sample X. Even though the density is obtained by applying the derivative operator to the distribution function, application of the same operator to the empirical distribution is meaningless and hence estimators have to be constructed otherwise.

1.3. Regression. Suppose we observe i.i.d. pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ sampled from an unknown joint distribution F on the product space $\mathcal{X}\times\mathcal{Y}$. In these pairs X_j's are thought of as explanatory variables and Y_j's are the corresponding responses. Typically in this setup it is required to estimate some functional of the conditional distribution of X_1 given Y_1. Below $\mathbb{P}_F$ denotes the probability induced by the data on the product space $(\mathcal{X}\times\mathcal{Y})^n$, and $\mathbb{E}_F$ stands for the corresponding expectation.

1.3.1. *Mean regression.* The problem of estimating the conditional mean

$$\mu_F(x)=\mathbb{E}_F(Y_1|X_1=x)$$

is known as the *mean regression*. Let $\varepsilon_j:=Y_j-\mu_F(X_j)$, then

$$Y_j=\mu_F(X_j)+\varepsilon_j,\quad j=1,\ldots,n. \tag{1.2}$$

Obviously, ε_j's have zero means, are i.i.d. by assumption, and $\mu(X_j)$ and ε_j are uncorrelated. Hence it is natural to consider the mean regression problem in a somewhat different format, assuming that the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ is generated by equation

$$Y_j=f(X_j)+\varepsilon_j,\quad j=1,\ldots,n, \tag{1.3}$$

where $f:\mathcal{X}\mapsto\mathcal{Y}$ is the unknown *regression function* and ε_j's are i.i.d. zero mean random variables, with either known or unknown distribution. The *design* points X_j can be either random or deterministic. An estimator of f is a function $\widehat{f_n}$ constructed using the available sample.

1.3.2. *Quantile regression.* The problem of estimating conditional p-quantile

$$Q_F(x;p)=\inf\Big\{y\in\mathbb{R}:\mathbb{P}_F(Y_1\le y|X_1=x)\ge p\Big\}$$

is known as the *quantile regression*. The special case $p=\frac{1}{2}$ corresponds to median.

1.3.3. *Classification.* When $\mathcal{Y}$ is finite, the problem of estimating conditional probabilities

$$p_F(y;x)=\mathbb{P}_F(Y_1=y|X_1=x),\quad y\in\mathcal{Y},$$

is called *classification*, and it is a fundamental problem in Machine Learning. In this context, points in $\mathcal{Y}$ are called *labels*, the space $\mathcal{X}$ is finite dimensional, typically $\mathbb{R}^d$, and the entries of $x\in\mathcal{X}$ are called *features*.

1.4. Normal means model. A basic problem in classical statistics is that of estimating the mean vector $\mu \in \mathbb{R}^d$ of a multivariate standard normal distribution from the sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, I_d)$. The sufficient statistic for this model is $\frac{1}{n}\sum_{j=1}^n X_j$, which has $N(\mu, n^{-1}I_d)$ distribution. Thus in a way, it is equivalent to the model $X \sim N(\mu, \varepsilon I_d)$, where $\varepsilon = 1/n$ can be viewed as the noise intensity. The nonparametric analog of this problem is estimation of the infinite sequence $(\mu_j)_{j\in\mathbb{N}}$ given the samples

$$X_j = \mu_j + \sqrt{\varepsilon} Z_j, \quad j = 1, 2, \ldots \tag{1.4}$$

where $Z_j \overset{\text{i.i.d.}}{\sim} N(0,1)$.

In disguise, this is the problem of estimating a continuous time signal observed in white noise. The typical scenario consists of the *continuous time* stochastic process,

$$X(t) = \mu(t) + \sqrt{\varepsilon} Z(t), \quad t \in [0, T], \tag{1.5}$$

where $\mu(t)$ is an unknown deterministic function, referred to as the *signal*, and $Z(t)$ is a random process, which models the *observation noise* of intensity $\sqrt{\varepsilon} > 0$.

The interval length T is finite and $T = 1$ can be taken without loss of generality. The signal μ is to be estimated from the observed trajectory $X = \{X(t), t \in [0,1]\}$. Note that the sample here is a continuum of random variables, rather than a finite number as in the previous problems. An estimator $\widehat{\mu}$ is a function on $[0,1]$, constructed using the sample X.

This problem can be reformulated in a completely different and equivalent form under additional assumptions. Consider the $L_2([0,1])$ space of functions with the norm

$$\|f\|_2 = \left(\int_0^1 f(t)^2 dt\right)^{1/2}.$$

For such functions the scalar product

$$\langle f, g\rangle = \int_0^1 f(t)g(t)dt, \quad f, g \in L_2([0,1]),$$

is well defined and finite by the Cauchy-Schwarz inequality.

Endowed with a scalar product such as this, $L_2([0,1])$ becomes the Hilbert space. It is the infinite dimensional counterpart of the linear spaces, familiar from the classical Linear Algebra. In particular, one can find an infinite sequence of orthonormal functions $\phi_j(t)$, $j = 1, 2, \ldots$

$$\langle \phi_i, \phi_j\rangle = \mathbf{1}_{\{i=j\}},$$

such that any other element $f \in L_2([0,1])$ has the representation

$$f(t) = \sum_{j=1}^{\infty} \langle f, \phi_j\rangle \phi_j(t), \tag{1.6}$$

where the series converges in the norm. Such a collection of functions is called complete orthonormal basis. Bases can be constructed using trigonometric functions, polynomials, etc.

המרחב $L^2(X)$ כאשר $X \subset \mathbb{R}$ הוא אחד מהמרחבים הפונקציונליים ומרחבי הילברט החשובים ביותר.

Assuming that μ and the paths of Z belong to $L_2([0,1])$ and taking scalar product of (1.5) with ϕ_j from some basis, we arrive at (1.4), where $X_j = \langle X, \phi_j \rangle$, $\mu_j = \langle \mu, \phi_j \rangle$ and $Z_j = \langle Z, \phi_j \rangle$. In view of the correspondence between functions and their coefficients in expansion (1.6), the original problem of estimating the function μ given a trajectory of X reduces to that of estimating the sequence $(\mu_j)_{j\in\mathbb{N}}$ given the sequence $(X_j)_{j\in\mathbb{N}}$.

If the noise $Z(t)$ is a centred Gaussian process[1], then Z_j's are normal with zeros means. Moreover, if Z_j's are uncorrelated and therefore also independent, the process $Z(t)$ is called *white noise*. Very roughly it can be thought of as a process with zero correlation at any distinct times. It is a common building block in engineering models, even though its rigorous mathematical construction is quite involved on the technical level.

1.5. Methods and Theory. Let us take a closer look at the mean regression problem for definiteness. It is obvious that without some smoothness assumption on the function f, nothing can be said about its values at any point distinct from all X_j's, even in absence of noise, i.e., when $\varepsilon_j = 0$. On the other hand, if f is smooth to a certain degree, e.g. differentiable a number of times, its values at the points in a small vicinity of X_j must be close to $f(X_j)$. Hence even noisy observations at the design points bare some local information about the values the unknown function.

This heuristics lies in foundations of all estimation methods, which, in one way or another, fit a smooth function to the data points. The accuracy of estimation is then affected by the following two sources of error:

(1) the *bias term* due to the approximation error of the unknown function by functions from a chosen class, and

(2) the *variance term* of the random deviations induced by the noise.

Assuming that the actual unknown function belongs to a class of functions with a certain degree of smoothness, the estimator can be tuned by finding the optimal balance between these two error contributions.

The quality of estimators is assessed within the usual decision theoretic framework. Let ρ be a metric on the space, to which the unknown function is assumed to belong, and define the *risk* functional

$$R(f,\widehat{f}) := \mathbb{E}_f w\big(\rho(f,\widehat{f})\big),$$

where w is some sufficiently regular nondecreasing function. As in the parametric statistics, it is typically impossible to prefer estimators on the basis of their risks directly, and either minimax or Bayes approaches are to be taken.

[1] A process $X(t)$, $t \in [0,1]$ is Gaussian if the random variables $X(t_1),\ldots,X(t_n)$ are jointly normal for any finite collection of times $t_1 < \ldots < t_n$, $n \in \mathbb{N}$.

It is common to use metrics induced by norms and some standard risks are

$$\mathbb{E}_f\big(f(x_0)-\widehat{f}(x_0)\big)^2, \qquad \text{(Mean Squared Error)},$$
$$\mathbb{E}_f\int_I\big(f(x)-\widehat{f}(x)\big)^2dx, \qquad \text{(Mean Integrated Squared Error)},$$
$$\mathbb{E}_f\sup_{x\in I}\big|f(x)-\widehat{f}(x)\big|, \qquad \text{(Mean Uniform Absolute Error)}.$$

where I is the domain of f and x_0 is a fixed point in I.

2. Set estimation

The goal of set estimation is to use the available sample to construct a *confidence* set $C(X)$, which covers the true unknown value of the parameter

$$\mathbb{P}_f\big(f\in C(X)\big)\geq 1-\alpha, \quad f\in\forall\Theta,$$

for a given coverage probability $1-\alpha$. In the parametric setting, confidence sets are usually intervals in one dimension and various meaningful geometric forms, such as ellipsoids, cubes, etc., in higher dimension.

In nonparametric problems, confidence sets are often chosen as balls around a point estimator $\widehat{f}$, with respect to some norm $\|\cdot\|$,

$$C=\big\{f:\|f-\widehat{f}\|\leq r\big\}=:B_r(\widehat{f}).$$

An alternative is a confidence band,

$$C=\big\{f:L(x)\leq f(x)\leq U(x),\ \forall x\in I\big\},$$

defined by the upper and lower envelopes $U(x)$ and $L(x)$, constructed using the sample X.

3. Hypothesis testing

The general hypothesis testing problem is to decide whether the unknown function (distribution, density, regression, etc.), which determines the distribution of the sample, belongs to a specific subset $\mathcal{F}_0$ of an appropriate function space. In other words, based on a sample $X\sim\mathbb{P}_f$, it is required to decide whether the null hypothesis

$$H_0: f\in\mathcal{F}_0$$

is true. The complement of $\mathcal{F}_0$ in the relevant space form the subspace of *alternatives*. A function $\delta:\mathcal{X}\mapsto\{0,1\}$ is called *test*, and H_0 is rejected if and only if $\{\delta(X)=1\}$. Typically tests are sought in the form $\delta(X)=\{T(X)\geq c\}$, where $T:\mathcal{X}\mapsto\mathbb{R}$ is some *test statistic* and $c\in\mathbb{R}$ is the *critical value*.

When a test statistic is chosen, the critical value is determined by the significance level (or size) requirement

$$\sup_{f\in\mathcal{F}_0}\mathbb{P}_f(T(X)\geq c)\leq\alpha, \tag{3.1}$$

where α is a given small error probability. A good test is the one which has large power $\mathbb{P}_f(\delta(X)=1)$ at the alternatives of interest $f\notin\mathcal{F}_0$.

3.1. Goodness of fit problem. Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} F$ and let $\mathcal{F}_0$ be a certain given subset of distributions. The problem of testing the hypothesis

$$H_0 : F \in \mathcal{F}_0,$$

is called the goodness-of-fit problem. For example, if $\mathcal{F}_0$ consists of all normal distributions,

$$\mathcal{F}_0 = \left\{ \Phi\Big(\frac{\cdot - \mu}{\sigma}\Big) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+ \right\}$$

the hypothesis in question is whether the sample was drawn from a normal density.

3.2. One sample problems. Given a sample $X \sim \mathbb{P}_F$, it is often required to test for a particular property of F or some functional of it. For example, a problem which frequently arises in medical trials, is to test for zero mean

$$H_0 : \mu_F = 0.$$

Other examples are to test for monotonicity of the function in the regression problem, unimodality of the density, from which the sample has been drawn, etc.

3.3. Two sample problems. Were two given samples drawn from the same distribution? This question can be formalised as testing for equality

$$H_0 : F = G,$$

given two samples

$$X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} F \quad \text{and} \quad Y_1,\dots,Y_m \overset{\text{i.i.d.}}{\sim} G.$$

On can also ask a weaker question of whether F and G have the same mean,

$$H_0 : \mu_F = \mu_G,$$

or compare any other particular feature of importance.

3.4. Independence tests. A sample of pairs $(X_1,Y_1),\dots,(X_n,Y_n) \overset{\text{i.i.d.}}{\sim} F$ is observed and it is required to test whether the entries within the pairs are independent, that is, whether the joint c.d.f. factors into the product of its marginals,

$$H_0 : F(x,y) = F(x,\infty)F(\infty,y).$$

DEPARTMENT OF STATISTICS, THE HEBREW UNIVERSITY, MOUNT SCOPUS, JERUSALEM 91905, ISRAEL

E-mail address: Pavel.Chigansky@mail.huji.ac.il