Estimation of distribution function

(notes by Pavel Chigansky)

A basic problem in nonparametric statistics is estimation of the unknown distribution F function on $\mathcal{X} \subseteq \mathbb{R}^d$ from the sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$. We will denote by $\mathbb{P}_F$ the product probability on the space of sequences $\mathcal{X}^\infty$. Below we consider only the simpler one-dimensional case $d = 1$.

1. Empirical distribution

A reasonable estimator of F is the *empirical* distribution function

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

This piecewise constant function with jumps at the data points X_j's clearly qualifies as a legitimate (purely discrete) c.d.f.

1.1. Elementary properties. A simple calculation shows that $\widehat{F}_n$ is an *unbiased* estimator for F,

$$\mathbb{E}_F \widehat{F}_n(x) = F(x), \quad x \in \mathbb{R}. \tag{1.1}$$

Its MSE risk at any fixed point $x_0 \in \mathbb{R}$ satisfies the bound

$$\mathbb{E}_F \Big(\sqrt{n} \big(\widehat{F}_n(x_0) - F(x_0) \big) \Big)^2 = \mathrm{Var}_F \big(\sqrt{n} \widehat{F}_n(x_0) \big) = F(x_0)(1 - F(x_0)) \le \frac{1}{4},$$

uniformly over F. Consequently, $\widehat{F}_n$ is uniformly consistent at rate $\sqrt{n}$,

$$\lim_n \sup_F \mathbb{E}_F \big(\sqrt{n} \big(\widehat{F}_n(x_0) - F(x_0) \big) \big)^2 < \infty. \tag{1.2}$$

By the CLT,

$$\sqrt{n} \big(\widehat{F}_n(x_0) - F(x_0) \big) \xrightarrow[n \to \infty]{d(\mathbb{P}_F)} N\big(0, F(x_0)(1 - F(x_0))\big)$$

and consequently, the interval with endpoints at

$$\widehat{F}_n(x_0) \pm \frac{1}{\sqrt{n}} \sqrt{\widehat{F}_n(x_0)(1 - \widehat{F}_n(x_0))} z_{1-\alpha/2}$$

is an *asymptotic* confidence interval for $F(x_0)$ with coverage probability of $1 - \alpha$.

lecture notes for "Advanced Statistical Models B" course.

זה לא מדויק מה שכתוב כאן
להסתכל על שיעור 2

Similarly, an asymptotic confidence ellipsoid can be constructed for

$$F(x^k) = \big(F(x_1), \dots, F(x_k)\big)$$

at any given vector of points $x^k = (x_1, \dots, x_k)$ with $x_1 < \dots < x_k$. For any $x, y \in \mathbb{R}$,

$$Cov(\hat{F}_n(\overrightarrow{X^k}), \hat{F}_n(\overrightarrow{Y^k})) = = \frac{1}{n}\mathrm{Cov}_F\left(\widehat{F}_1(x), \widehat{F}_1(y)\right) = \frac{1}{n}\big(F(x \wedge y) - F(x)F(y)\big),$$

and hence by the multivariate CLT

$$\sqrt{n}\big(\widehat{F}_n(x^k) - F(x^k)\big) \xrightarrow[n\to\infty]{d(\mathbb{P}_F)} N\big(0, \Sigma_F(x^k)\big), \tag{1.3}$$

where $\Sigma_F(x^k)$ is the covariance matrix with the entries $F(x_i \wedge x_j) - F(x_i)F(x_j)$. It then follows that [1] for any F,

$$\left\|\sqrt{n}\big(\widehat{F}_n(x^k) - F(x^k)\big)\right\|^2_{\Sigma^{-1}_{\widehat{F}_n}} \xrightarrow[n\to\infty]{d(\mathbb{P}_F)} \chi^2_k$$

and consequently, the ellipsoid

$$E_\alpha(X^n) = \left\{v \in \mathbb{R}^k : \left\|\widehat{F}_n(x^k) - v\right\|^2_{\Sigma^{-1}_{\widehat{F}_n}} \le \frac{1}{n}\chi^2_{k,1-\alpha}\right\}$$

is an asymptotic confidence set for $F(x^k)$ with coverage probability of $1-\alpha$. Some additional elementary properties are explored in Problem 1.

1.2. Consistency with respect to supremum norm. The elementary results, derived in the previous subsection, are relevant, when we are interested in estimating the value of F at a single given point x_0 or, more generally, at any set of finite points. What about estimation of F at all points of $\mathbb{R}$ *simultaneously*?

The first and basic question in this regard is whether $\widehat{F}_n$ estimates F consistently with respect to a given metric ρ on the space of distribution functions, that is,

$$\rho(\widehat{F}_n, F) \xrightarrow[n\to\infty]{\mathbb{P}_F} 0,$$

for any F or, perhaps, uniformly over all F or some class of F's? An affirmative answer can be given for some natural metrics by elementary means (see Problems 2 and 3).

A particularly useful metric is the one defined by the supremum norm,

$$\|\widehat{F}_n - F\|_\infty = \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right|.$$

This norm is important, since it can be used to construct confidence bands for F, as we will shortly see.

THEOREM 1.1 (Glivenko-Cantelli).

$$\|\widehat{F}_n - F\|_\infty \xrightarrow[n\to\infty]{\mathbb{P}_F - a.s.} 0, \quad \forall F.$$

[1] for $x \in \mathbb{R}^k$ and a positive definite matrix Σ, the function $\|x\|^2_\Sigma = x^\top \Sigma x$ is a norm

PROOF. Fix an integer m and define a nondecreasing sequence of points

$$x_j = \min\{x \in \mathbb{R} : F(x) \geq j/m\}, \quad j = 1, \dots, m-1,$$

and $x_0 = -\infty$ and $x_m = \infty$. This definition is correct, since F is right continuous and therefor the minimum is attained. Note that if F is discrete, it is possible that some x_j's are equal (think of e.g. Ber$(1/2)$ distribution). If $x_{j+1} > x_j$, by this definition,

$$F(x_{j+1}-) - F(x_j) \leq \frac{j+1}{m} - \frac{j}{m} \leq 1/m.$$

Then for all j such that $x_j < x_{j+1}$ and any $x \in [x_j, x_{j+1})$,

$$\widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_{j+1}-) - F(x_j) \leq \widehat{F}_n(x_{j+1}-) - F(x_{j+1}-) + 1/m,$$
$$\widehat{F}_n(x) - F(x) \geq \widehat{F}_n(x_j) - F(x_{j+1}-) \geq \widehat{F}_n(x_j) - F(x_j) - 1/m.$$

Thus

$$\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \leq \max_{0\leq j\leq m} \left(\left|\widehat{F}_n(x_j-) - F(x_j-)\right| \vee \left|\widehat{F}_n(x_j) - F(x_j)\right| \right) + 1/m.$$

The expression in the right hand side converges[2] to zero $\mathbb{P}_F$-a.s. as $n \to \infty$ by the strong LLN, since maximum is taken over a finite number of terms. The claim follows by arbitrariness of m.

□

2. Confidence bands

Consider the random variable

$$D_n := \sqrt{n} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)|.$$

If F has atoms, the distribution of D_n may depend on F, both for fixed n and in the limit, see Problem 4. However, when F is continuous (Problem 5),

$$\big(F(X_1), \dots, F(X_n)\big) \stackrel{d}{=} \big(V_1, \dots, V_n\big), \tag{2.1}$$

where $V_j \stackrel{\text{i.i.d.}}{\sim} U([0,1])$ and we can write

$$D_n = \sqrt{n} \sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\{F(X_j)\leq F(x)\}} - F(x) \right| = \sqrt{n} \sup_{u\in[0,1]} \left| \frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\{V_j \leq u\}} - u \right|.$$

The important implication is that D_n is *distribution free*: its distribution under $\mathbb{P}_F$ does not depend on F. This property is central to construction of confidence bands. Indeed, let us denote by $\kappa_{n,p}$ the p-the quantile of the distribution of D_n and define the upper and lower *confidence* envelopes

$$U_n(x) = \widehat{F}_n(x) + \frac{1}{\sqrt{n}} \kappa_{n,1-\alpha} \quad \text{and} \quad L_n(x) = \widehat{F}_n(x) - \frac{1}{\sqrt{n}} \kappa_{n,1-\alpha}. \tag{2.2}$$

Then the set statistic

$$C_n(X^n) = \{F : L_n(x) \leq F(x) \leq U_n(x), \quad \forall x \in \mathbb{R}\}$$

[2] Note that $\widehat{F}_n(x-) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{X_j < x\}}$.

is an *exact* confidence band with coverage probability $1-\alpha$,

$$\mathbb{P}_F\big(F \in C_n(X^n)\big) = \mathbb{P}_F\big(D_n \le \kappa_{n,1-\alpha}\big) = 1-\alpha, \quad \forall F.$$

2.1. Asymptotic confidence bands. While, at least in principle, quantiles of D_n can be found in an explicit form, their actual computation becomes very complicated beyond unrealistically small sample sizes (see Problem 6). Does the distribution of D_n converge when $n \to \infty$ and if so, what is the limit? Such a limit would be useful for approximation of $\kappa_{n,1-\alpha}$ for large n.

THEOREM 2.1 (A.Kolmogorov [2]). *Assume F is continuous, then*

$$\mathbb{P}_F(D_n \le x) \xrightarrow[n\to\infty]{} \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}, \quad x \ge 0.$$

The series cannot be summed in a closed form, but its terms decay very quickly with j and the quantile of Kolmogorov's distribution $\kappa_{1-\alpha}$ is not hard to approximate numerically within any desired precision (implementations are routinely available in common statistical software packages). Consequently by replacing $\kappa_{n,1-\alpha}$ with $\kappa_{1-\alpha}$ in the confidence envelopes (2.2), an *asymptotic* confidence band with coverage probability of $1-\alpha$ is obtained,

$$\mathbb{P}_F\big(F \in C_n(X^n)\big) \xrightarrow[n\to\infty]{} 1-\alpha, \quad \forall F.$$

2.2. Donsker's theorem. The original proof of Theorem 2.1 in [2] involves combinatorial arguments and requires basic knowledge of partial differential equations. An alternative approach, based on the theory of random processes, is much more involved on the technical level, but provides an interesting insight.

Since D_n is distribution-free, only the uniform distribution on $[0,1]$ with $F(x) = x$ can be considered without loss of generality. The key idea then is to consider the sequence of stochastic processes

$$\overline{B}_n(x) := \sqrt{n}\big(\widehat{F}_n(x) - x\big), \quad x \in [0,1], \tag{2.3}$$

where $\widehat{F}_n(x)$ is the empirical distribution of $X_1,\dots,X_n \sim U([0,1])$. In view of (1.3), for any $k \in \mathbb{N}$,

$$\sqrt{n}\big(\widehat{F}_n(x) - x\big) \xrightarrow[n\to\infty]{d} N\big(0,\Sigma(x)\big), \quad x = (x_1,\dots,x_k) \in \mathbb{R}^k,$$

where $\widehat{F}_n(x)$ is the vector with entries $\widehat{F}_n(x_j)$ and the covariance matrix $\Sigma(x)$ is given by

$$\Sigma_{ij}(x) = x_i \wedge x_j - x_i x_j.$$

Hence if the sequence of processes $\overline{B}_n(\cdot)$ converges in a meaningful sense, the limit $\overline{B}(\cdot)$ must be a Gaussian process[3] with zero mean $\mathbb{E}\overline{B}(x) = 0$ and covariance function

$$\mathbb{E}\overline{B}(x)\overline{B}(y) = x \wedge y - xy, \quad x,y \in [0,1]. \tag{2.4}$$

It then makes sense to expect that the Kolmogorov distribution must coincide with the distribution of the random variable $\sup_{x\in[0,1]} \overline{B}(x)$. Implementation of this program poses the following questions.

(Q1) Does there exist a Gaussian process $\overline{B}(\cdot)$ with zero mean and covariance function as above?

(Q2) How to define and prove *convergence in distribution* for processes

$$\overline{B}_n(\cdot) \xrightarrow[n\to\infty]{d} \overline{B}(\cdot),$$

so that it implies convergence in distribution of functionals such as,

$$\sup_{x\in[0,1]} |\overline{B}_n(x)| \xrightarrow[n\to\infty]{d} \sup_{x\in[0,1]} |\overline{B}(x)| \ ?$$

(Q3) How to find the distribution of $\sup_{x\in[0,1]} |\overline{B}(x)|$?

Adequate answers to these questions require sophisticated mathematics. Construction of sufficiently regular processes with given finite dimensional distributions is a fundamental problem in the theory of random processes, essentially solved by A.Kolmogorov himself. The limit process $\overline{B}(\cdot)$ in question (Q1), called the *Brownian bridge*, turns out to have continuous but nowhere differentiable paths, all of which start and terminate at zero at the endpoints of the interval $[0,1]$, see an illustration at Figure 1. The appropriate notion of convergence in (Q2), called the *weak* convergence, is the one implied by convergence of expectations of all bounded functionals, continuous in the relevant topology, in this case, the one induced by the supremum norm. Finally, the distribution in (Q3) does indeed coincide with that in Theorem 2.1, as can be shown by means of tools from *stochastic calculus*. All this is summarized by the following famous result, which is a basic example of a *functional* CLT.

THEOREM 2.2 (M.Donsker). *For any continuous distribution F,*

$$\Big(\sqrt{n}\big(\widehat{F}_n(x) - F(x)\big),\ x\in\mathbb{R}\Big) \xrightarrow[n\to\infty]{w} \Big(\overline{B}(F(x)),\ x\in\mathbb{R}\Big),$$

where $\overline{B}(\cdot)$ is the Brownian bridge, i.e. the unique Gaussian random process with zero mean and covariance (2.4).

2.3. Nonasymptotic confidence bands. The confidence band with the envelopes (2.2) is exact, but the actual calculation of $\kappa_{n,1-\alpha}$ for a given n is infeasible. If n is large, then using Kolmogorov's $\kappa_{1-\alpha}$ instead gives a confidence band, whose actual coverage for any fixed n will be only close to $1-\alpha$ and the approximation would improve with $n\to\infty$. Yet, quite remarkably, an exact (non-asymptotic) confidence band can be constructed, using the following large deviations type result.

[3]A random process $X(\cdot)$ is Gaussian if its finite dimensional distributions, i.e. distributions of all vectors of the form

$$\big(X(t_1),\ldots,X(t_k)\big), \quad t\in\mathbb{R}^k, \quad k\in\mathbb{N},$$

are Gaussian.

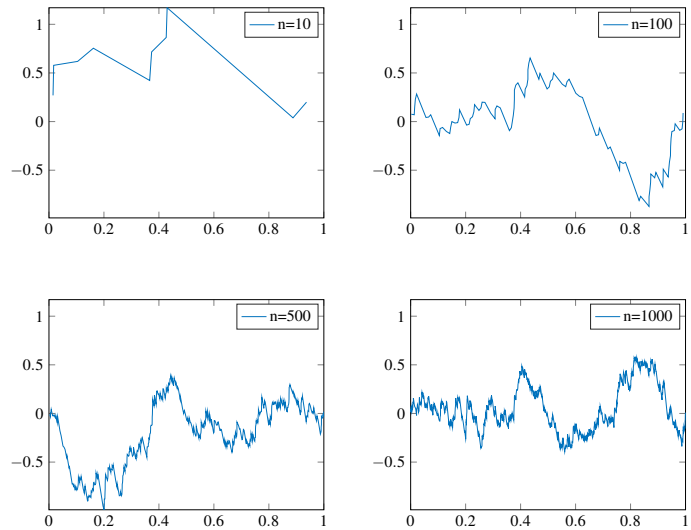

FIGURE 1. Evolution of a typical trajectory of

$$\overline{B}_n(x) = \sqrt{n}(\widehat{F}_n(x) - x)$$

is depicted as n increases. The last picture indicates how a typical trajectory of the limit process, i.e. the Brownian bridge, may look like: a continuous but highly irregular function.

THEOREM 2.3 (Dvoretzky, Kiefer & Wolfowitz). *For any continuous F,*

$$\mathbb{P}_F\Big(\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \geq \varepsilon\Big) \leq 2e^{-2n\varepsilon^2}, \quad \forall \varepsilon > 0.$$

This inequality bounds the tails of the distribution for any fixed sample size n and any $\varepsilon > 0$. Hence we can take $\varepsilon := c_\alpha/\sqrt{n}$ with $c_\alpha = \sqrt{\frac{1}{2}\log\frac{2}{\alpha}}$ and

$$U_n(x) = \widehat{F}_n(x) + \frac{1}{\sqrt{n}}c_\alpha \quad \text{and} \quad L_n(x) = \widehat{F}_n(x) - \frac{1}{\sqrt{n}}c_\alpha, \tag{2.5}$$

to obtain non-asymptotic confidence band

$$\mathbb{P}_F\Big(L_n(x) \leq F(x) \leq U_n(x),\ \forall x \in \mathbb{R}\Big) \geq 1 - \alpha, \quad \forall F \in C(\mathbb{R}).$$

Numerically the values of Kolmogorov's $\kappa_{1-\alpha}$ and c_α in (2.5) are very close, for example at $\alpha = 0.05$ the relative error is only 0.00021... percent. On one hand, this implies that, for all practical purposes, the asymptotic band based on the Kolmogorov distribution may only be conservative for small n and, on the other hand, since it is asymptotically exact, the DKW bound is quite tight.

3. Optimality

As we saw above, the empirical distribution estimator is consistent at rate $\sqrt{n}$ with respect to several risks, uniformly over *all* distribution functions. Are there

estimators with a better rate? As in parametric statistics, this question is meaningless, if not asked properly: the useless estimator $\widehat{F} = F_0$ is capable of estimating one particular distribution, namely F_0, *exactly*, even disregarding the sample.

The more meaningful question is whether estimation at a better rate is possible uniformly over all distribution functions. A negative answer to this question was given in [1],

$$\lim_n \sup_F \mathbb{E}_F \|\sqrt{n}(\widehat{F}_n - F)\| = \lim_n \inf_{\widetilde{F}_n} \sup_F \mathbb{E}_F \|\sqrt{n}(\widetilde{F}_n - F)\|$$

for several norms, including $\|\cdot\|_\infty$, where the infimum is taken over all estimators based on a sample of size n. This implies that $\widehat{F}_n$ is asymptotically optimal in the minimax sense.

Computer experiment

Implement computation of the confidence bands based on Theorem 2.1 and Theorem 2.3. For $n = 100$ and some true continuous distribution, estimate the actual coverage probabilities by the proportions of coverages in $M = 10,000$ Monte Carlo trials. Compare your results with the theory presented in this chapter.

Exercises

PROBLEM 1. This problem explores some additional elementary properties of the empirical distribution.

(1) Argue that under $\mathbb{P}_F$,

$$n\widehat{F}_n(x) \sim \mathrm{Bin}(n, F(x)).$$

(2) Prove that for any $x_1 < \ldots < x_k$, the vector of increments

$$n\big(\widehat{F}_n(x_1), \widehat{F}_n(x_2) - \widehat{F}_n(x_1), \ldots, \widehat{F}_n(x_k) - \widehat{F}_n(x_{k-1}), 1 - \widehat{F}_n(x_k)\big)$$

has multinomial distribution and find its parameters.

PROBLEM 2. Argue that for any distribution F

$$\mathbb{E}_F \int_{\mathbb{R}} \Big(\sqrt{n}(\widehat{F}_n(x) - F(x))\Big)^2 dF(x) \le \frac{1}{4},$$

and show that, when F is continuous, this integral, in fact, equals $1/6$. Deduce consistency with respect to the respective metric.

PROBLEM 3.

1) Show that the MISE satisfies

$$\sup_F \mathbb{E}_F \big\|\sqrt{n}(\widehat{F}_n - F)\big\|_2^2 = \infty,$$

where the supremum is taken over all distributions.

2) Show that the MISE satisfies

$$\sup_{F \in \mathcal{F}_M} \mathbb{E}_F \big\|\sqrt{n}(\widehat{F}_n - F)\big\|_2^2 \le M,$$

where $\mathcal{F}_M = \{F : \int_{\mathbb{R}} |x| dF(x) \le M\}$. Deduce uniform consistency on $\mathcal{F}_M$ with respect to $L^2(\mathbb{R})$ norm.

Hint: for a random variable $\xi \ge 0$,

$$\mathbb{E}\xi = \int_0^\infty \big(1 - F_\xi(x)\big) dx.$$

PROBLEM 4. This problem explores properties of

$$D_n = \sqrt{n} \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)|$$

when F is purely discrete with atoms at $\{x_1, ..., x_k\}$ and probabilities $\{p_1, ..., p_k\}$.

(1) Does the distribution of D_n depend on x_j or/and on p_j's, when n is fixed?

(2) Does distribution of D_n converge as $n \to \infty$? Does it depend on F?

PROBLEM 5. Prove (2.1).

PROBLEM 6. This problem explores some properties of D_n for continuous distributions.

(1) Derive the formula[4]

$$D_n = \sqrt{n} \max_{j\le n} \left(\left| \frac{j}{n} - F(X_{(j)}) \right| \vee \left| \frac{j-1}{n} - F(X_{(j)}) \right| \right),$$

where $X_{(j)}$ is the j-th order statistic.

(2) Find the probability density of D_n for $n = 2$.

PROBLEM 7. Suppose $X_1, ..., X_n$ are i.i.d. samples from an unknown continuous distribution F. If no a priori knowledge on F is available, the empirical distribution can be used to estimate F and the corresponding normalized MISE risk was found in Problem 2 to be equal to $1/6$. Suppose now that it is revealed that F belongs to the parametric family $(F_\lambda)_{\lambda\in\mathbb{R}_+}$ where F_λ is $\mathrm{Exp}(\lambda)$ distribution. Compute the large sample limit of the normalized MISE risk of the estimator $F_{\widehat{\lambda}_n}$ where $\widehat{\lambda}_n$ is the MLE of λ.

References

[1] Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. (1956), "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator", Annals of Mathematical Statistics, 27 (3): 642669

[2] Shiryayev, A. N. On The Empirical Determination of A Distribution Law. Selected Works of AN Kolmogorov. Springer, Dordrecht, 1992. 139-146.

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

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

[4] This formula is usually used to compute D_n