A modified problem by Ya. Yu. Nikitin

Anna Tchirina

Saint-Petersburg Electrotechnical University, Saint Petersburg, Russia

St. Petersburg, August 31, 2021
YAKOV NIKITIN MEMORIAL SESSION

Joint work with A.I. Nazarov, to appear in ZNS POMI

The problem setting

Let X_j be independent observations of a r.v. X having a continuous d.f. F. Consider the problem of testing the hypothesis $F = F_0$ (with F_0 a fully specified d.f.) against the alternative $F \neq F_0$.

One of the classical methods to construct nonparametric goodness-of-fit tests is using various functionals $\mathcal{J}[\xi_n]$ of the (transformed) empirical process

$$\xi_n(s) := \sqrt{n}\Big(n^{-1}\sum_{j=1}^{n} \mathbf{1}\{F_0(X_j) \le s\} - s\Big). \tag{1}$$

To compare different test statistics we consider the so-called *Bahadur exact slope*, that is a non-random positive function of the parameter θ which describes the alternative hypothesis. See Nikitin, 1995, §1.2.

The problem setting

Since computing the exact slope is quite a complicated problem, it is often restricted to the study of the *local* exact slope, i.e. the asymptotics of the exact slope as $F \to F_0$. Under quite general assumptions this asymptotics has the form $b\theta^2$ as $\theta \to 0$, the coefficient b is called *local Bahadur index* of the corresponding sequence of test statistics. See Nikitin, 1995, Chapter II.

It follows from the Bahadur–Raghavachari Theorem that for any sequence of statistics the exact Bahadur slope does not exceed twice the Kullback – Leibler information. So, the statistics for which the ratio of these two quantities at least tends to 1 as $F \to F_0$, are of particular interest. They are called *locally asymptotically optimal in the Bahadur sense* (LAO).

In other words, the LAO statistics have the maximum possible local Bahadur index for a given family of alternatives.

The problem setting

For some concrete families of alternatives, such as shift, scale etc., the LAO property of the sequence of test statistics is determined by the null hypothesis F_0.

Ya.Yu. Nikitin posed an inverse problem:

To find the distributions for which a given sequence of statistics is LAO.

This problem was discussed in details in Nikitin, 1995, §§6.2-6.3, and was solved in many cases. However, for some statistics the set of such distributions turns out to be empty. Let's explain it.

The problem setting

Under some regularity conditions, the corresponding d.f. F_0 is a solution of the differential equation

$$\frac{\partial}{\partial\theta} F_\theta(x)\big|_{\theta=0} = \mathfrak{u}(F_0(x)), \tag{2}$$

with $\mathfrak{u}$ one of the *leading functions* for the sequence of statistics. This means that the function $\mathfrak{u}$ maximizes the functional $\mathcal{J}[u]$ on the unit ball in the Sobolev space $\overset{\circ}{W}{}_2^1(0,1)$. See Nikitin, 1995, §§6.1 and 2.6.

However for a family with a shift parameter $F_\theta(x) = F(x+\theta)$ the equation (2) can be rewritten as follows:

$$F_0'(x) = \mathfrak{u}(F_0(x)),$$

which yields that the function $\mathfrak{u}$ must be **positive** on the interval $(0,1)$. This requirement holds as well for some other families such as scale families having the support on the positive half-line.

The problem setting

For many sequences of test statistics, including classic tests of Kolmogorov, Watson – Darling, Cramer – von Mises, Anderson – Darling etc., the set of leading functions contains a function positive on $(0,1)$, see, e.g., Nikitin, 1995, §6.2. However it is not true for some more sophisticated statistics.

In such a situation a natural question arises: what is the maximal *available* local asymptotic efficiency of such statistics (say, under the shift alternative).

To answer this question, we have to solve a modified problem: to maximize the corresponding functional $\mathcal{J}[u]$ on the set of *non-negative* functions from the unit ball in $\overset{\circ}{W}{}_2^1(0,1)$.

Notice that the maximizer necessarily vanishes in the interior points of the interval $[0,1]$ and hence it corresponds to a distribution with a discontinuous d.f., which is impossible under the original assumptions. Nevertheless, this maximizer can be approximated by functions *positive on* $(0,1)$. Therefore, though the found LAO still is not attained, for any $\varepsilon > 0$ there exist distributions for which it is attained up to ε.

Two examples

1. The following integrated analogs of the Watson statistic were introduced in N. Henze and Ya.Yu. Nikitin (2002):

$$\tilde{U}_n^2 = \int_0^1 \Big(A_n(s) - \int_0^1 A_n(t)\,dt \Big)^2 ds; \quad \bar{U}_n^2 = \int_0^1 \big(A_n(s) - sA_n(1) \big)^2 ds,$$

where $A_n(s) = \int_0^s \xi_n(t)\,dt$ (the process $\xi_n(t)$ was defined in (1)).

The statistic $\tilde{U}_n^2$ was shown to be LAO (under the shift alternative) for the hyperbolic cosine distribution with density $(\pi \cosh(x))^{-1}$). At the same time for the statistic $\bar{U}_n^2$ the maximizer in the corresponding extremal problem

$$\begin{aligned} \mathcal{J}_1[u] &:= \int_0^1 \Big(\int_0^x u(t)\,dt - x \int_0^1 u(t)\,dt \Big)^2 dx \to \max; \\ \mathcal{I}[u] &:= \int_0^1 (u'(x))^2\,dx \le 1; \qquad u(0) = u(1) = 0 \end{aligned} \tag{3}$$

is sign-changing, and therefore this statistic is optimal for no distribution under the shift alternative.

Two examples

2. O.A. Podkorytova studied in her PhD thesis (1994), among others, the statistics $S_{p,n} = \|\nu_n\|^2_{L_p(0,1)}$ based on the Deheuvels transformation of the empirical process

$$\nu_n(s) = \xi_n(s) + \int_0^s \frac{\xi_n(r)}{1-r}\,dr - s\int_0^1 \frac{\xi_n(r)}{1-r}\,dr.$$

It was shown that in the corresponding extremal problem (if $p = \infty$, the integral in $\mathcal{J}_{2,p}[u]$ should be replaced by the maximum)

$$\begin{aligned} \mathcal{J}_{2,p}[u] &:= \left[\int_0^1 \left|u(x) + \int_0^x \frac{u(t)}{1-t}\,dt - x\int_0^1 \frac{u(t)}{1-t}\,dt\right|^p dx\right]^{\frac{2}{p}} \to \max; \\ \mathcal{I}[u] &= \int_0^1 (u'(x))^2\,dx \le 1; \qquad u(0) = u(1) = 0 \end{aligned} \tag{4}$$

the maximizing function is sign-changing, and these statistics are not LAO for the shift alternative under any distribution.

Statistic $\bar{U}_n^2$

We are searching for the maximum in the extremal problem (3) under the extra condition $u \geq 0$. By standard variational argument, the maximum in this problem is attained. Moreover, by the homogeneity of the functionals $\mathcal{I}$ and $\mathcal{J}_1$ we can consider an equivalent problem

$$\mathcal{P}_1[u] := \frac{\mathcal{J}_1[u]}{\mathcal{I}[u]} \to \max; \qquad u \geq 0; \quad u \not\equiv 0; \quad u(0) = u(1) = 0.$$

The necessary condition of maximum is (here $\lambda = \mathcal{P}_1^{-1}[u] > 0$)

$$\frac{1}{2}\, d(\lambda \mathcal{J}_1 - \mathcal{I})[u]\eta \equiv \langle u'', \eta \rangle + \lambda \int_0^1 f_1(x)\eta(x)\, dx \leq 0 \tag{5}$$

for all variations $\eta \in \overset{\circ}{W}{}_2^1(0,1)$ such that $u + \eta \geq 0$. Here u'' is understood in the sense of distributions, and

$$f_1(x) = \int_x^1 (U(t) - tU(1))\, dt - \int_0^1 (tU(t) - t^2 U(1))\, dt; \quad U(x) = \int_0^x u(t)\, dt.$$

The steps to solution

- We prove that $u'' + \lambda f_1 \leq 0$ in the sense of distributions. Hence u'' is a measure (a charge), and moreover its singular component is non-positive.
- A set where $u > 0$ is an at most countable union of intervals. At any of these intervals I we have in fact $-u'' = \lambda f_1$. Thus, the support of the singular component of the measure u'' is contained in a set where $u = 0$.
- We prove that u'' has no singular component at all. Therefore, u' is continuous on $[0, 1]$. Therefore, at the ends of each interval I, besides the condition $u = 0$, we also have $u' = 0$ (except maybe the points 0 and 1).
- We prove that there is only finite number of intervals I. The complement of the closure of these intervals consists of a finite number of intervals, on which $u'' = 0$ holds, and therefore u'' is piecewise continuous on $[0, 1]$. Note that this complement is non-empty. Otherwise the function u would maximize the functional $\mathcal{P}_1[u]$ without the restriction $u \geq 0$, which is impossible due to Henze and Nikitin.

The steps to solution (cont.)

- If there is an isolated segment $[x_0, x_1] \subset (0, 1)$ on which the equation $-u'' = \lambda f_1$ holds then we can move it by a sufficiently small number h, keeping the other segments in place. Direct calculation gives that $\mathcal{P}_1[u_h] > \mathcal{P}_1[u]$, which is impossible.
- Thus, only three variants remain:
  1. $-u'' = \lambda f_1$ on $[0, x_0]$ and $u = 0$ on $[x_0, 1]$;
  2. $-u'' = \lambda f_1$ on $[x_1, 1]$ and $u = 0$ on $[0, x_1]$;
  3. $-u'' = \lambda f_1$ on $[0, x_0]$ and on $[x_1, 1]$, and $u = 0$ on $[x_0, x_1]$.

  The variants 1 and 2 are equivalent by symmetry. Further, we show that the variant 3 is unprofitable.
- We differentiate the equation $-u'' = \lambda f_1$ twice and obtain

$$u^{IV}(x) = \lambda\Big(u(x) - \int_0^1 u(t)\,dt\Big).$$

Completion of the solution

$$u^{IV}(x) = \lambda\Big(u(x) - \int_0^1 u(t)\,dt\Big). \tag{6}$$

A general solution of (6) for positive λ is

$$u(x) = c_1\cosh(kx) + c_2\sinh(kx) + c_3\cos(kx) + c_4\sin(kx) + c_5, \qquad k^4 = \lambda.$$

This solution should satisfy two boundary conditions at zero, two boundary conditions at x_0 and two matching conditions. This gives the nonlinear system for parameters k and $x_0 \in (0,1)$:

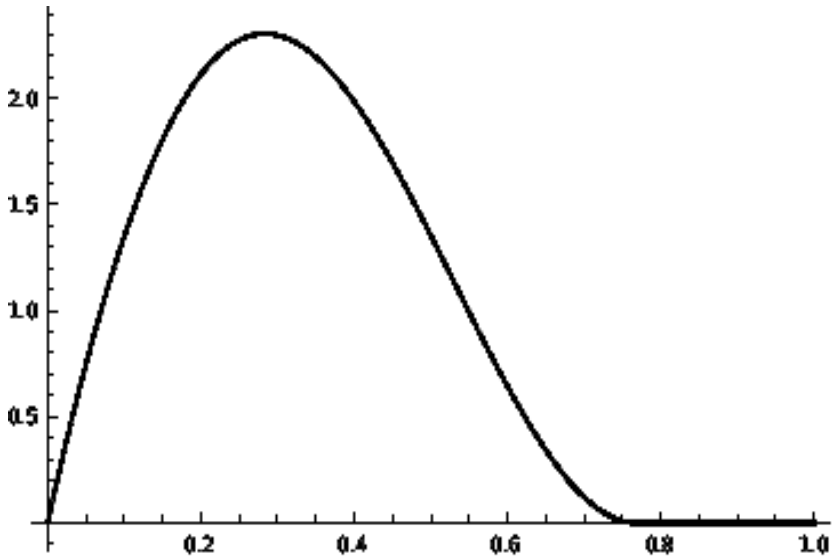
$$\sin\Big(\frac{kx_0}{2}\Big)\sinh\Big(\frac{kx_0}{2}\Big)\cdot\Big[\tanh\Big(\frac{kx_0}{2}\Big) + \tan\Big(\frac{kx_0}{2}\Big) + k(1-x_0)\Big] = 0. \tag{7}$$

$$\begin{aligned}&(6 - 3k^2x_0^2)\tan\Big(\frac{kx_0}{2}\Big) - (6 + 3k^2x_0^2)\tanh\Big(\frac{kx_0}{2}\Big)\\ &+ k^3(2 - 6x_0 + 3x_0^2 + x_0^3) - 6k(1-x_0)\tan\Big(\frac{kx_0}{2}\Big)\tanh\Big(\frac{kx_0}{2}\Big) = 0.\end{aligned} \tag{8}$$

The solution $u(x)$ is determined up to a multiplicative constant.

Completion of the solution

The system (7)–(8) has no solutions with $k \in (0, \pi)$ and has a unique solution with $k \in (\pi, 2\pi)$: $\widehat{k} \approx 5.21579$ and $\widehat{x}_0 \approx 0.767426$. The graph of the corresponding maximizer u is given on the figure.



Since $\lambda = \widehat{k}^4$ takes the minimal possible value, the function u maximizes the functional $\mathcal{P}_1$ under given restrictions, and the maximal value is $\widehat{k}^{-4} \approx 0.0013512$.

The optimal local index for $\bar{U}_n^2$, computed in Henze and Nikitin, equals ≈ 0.0019977. Thus the available local asymptotic efficiency of the statistic $\bar{U}_n^2$ under the shift alternative is more than 67% from the optimum.

Statistic $S_{\infty,n}$

We are interested in the maximum in the extremal problem

$$\mathcal{J}_{2,\infty}[u] := \max_{[0,1]} \left(u(x) + \int_0^x \frac{u(t)}{1-t}\, dt - x \int_0^1 \frac{u(t)}{1-t}\, dt \right)^2 \to \max;$$
$$\mathcal{I}[u] = \int_0^1 (u'(x))^2\, dx \le 1; \qquad u \ge 0; \quad u(0) = u(1) = 0. \tag{9}$$

Again, the standard variational argument shows the maximum in this problem is attained.
We proceed in two steps. First find the maximum of an auxiliary functional

$$\widetilde{\mathcal{J}}_{2,\infty}[u] = \left(u(x_*) + \int_0^{x_*} \frac{u(t)}{1-t}\, dt - x_* \int_0^1 \frac{u(t)}{1-t}\, dt \right)^2,$$

where $x_* \in (0,1)$ is fixed. Then maximize the result in $x_* \in (0,1)$.

Statistic $S_{\infty,n}$

The necessary condition of maximum is (here $\lambda = \mathcal{I}[u]/\widetilde{\mathcal{J}}_{2,\infty}[u] > 0$)

$$\frac{1}{2}\, d(\lambda\widetilde{\mathcal{J}}_{2,\infty} - \mathcal{I})[u]\eta \equiv \langle u'' + \lambda f_2, \eta\rangle \le 0 \tag{10}$$

for all variations $\eta \in \overset{\circ}{W}{}_2^1(0,1)$ such that $u + \eta \ge 0$. Here u'' is understood in the sense of distributions, and f_2 a measure (a charge)

$$f_2(x) = (v(x_*) - x_* v(1))\Big[\delta(x - x_*) + \frac{1}{1-x}\big(\chi_{[0,x_*]}(x) - x_*\big)\Big];$$

hereinafter

$$v(x) = u(x) + \int_0^x \frac{u(t)}{1-t}\, dt \tag{11}$$

is the Khmaladze transform of u.

The steps to solution

- As before, we obtain $u'' + \lambda f_2 \leq 0$ in the sense of distributions. Therefore u'' is a measure (a charge), and its singular component is non-positive. Next, on each interval I where $u > 0$ we have $-u'' = \lambda f_2$.
- As before, we deduce that $u'' + \lambda(v(x_*) - x_* v(1))\delta(\cdot - x_*)$ has no singular component. So, u' is continuous on $[0,1] \setminus \{x_*\}$, and, besides the condition $u = 0$, the condition $u' = 0$ holds on every interval I, except maybe for 0 and 1.
- We prove that there is a unique interval I. So, we have $-u'' = \lambda f_2$ on $[0, x_0]$ and $u = 0$ on $[x_0, 1]$.

  Moreover, $0 < x_* < x_0 < 1$, as otherwise the function u would maximize the functional without the restriction $u \geq 0$, which is impossible due to Podkorytova.

The steps to solution (cont.)

- We solve the differential equations on $[0, x_0]$ subject to the boundary conditions and matching conditions. This gives a nonlinear equation for x_* and x_0:

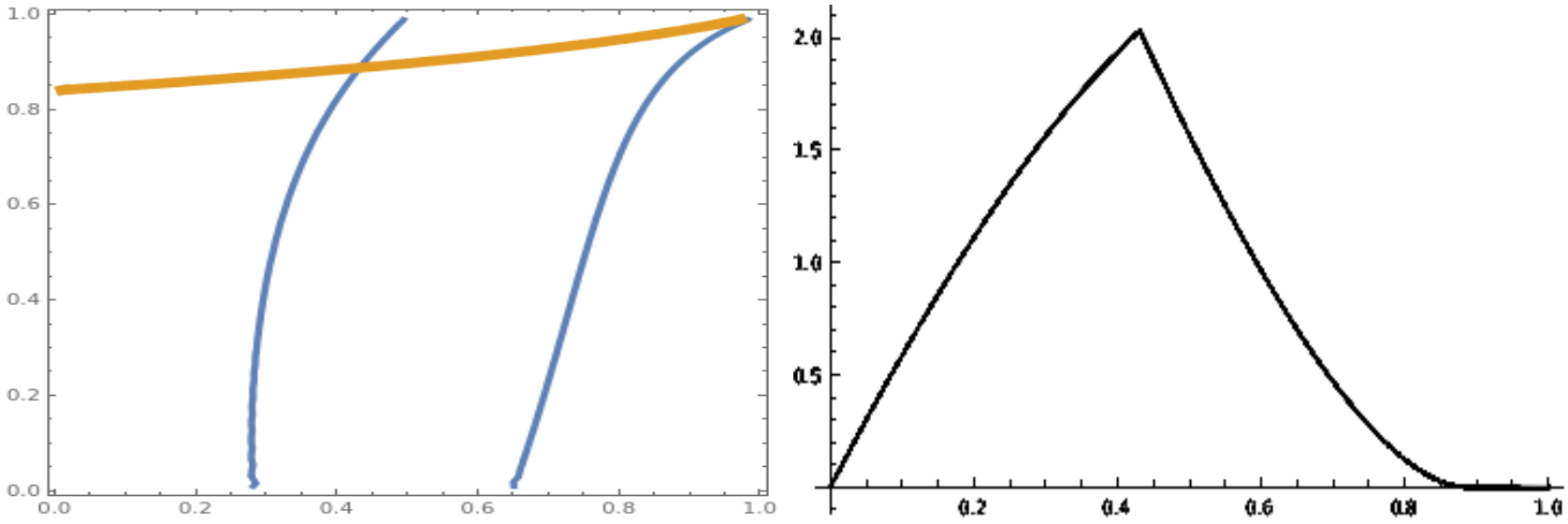
$$G(x_*, x_0) \equiv x_0 + \ln(1 - x_0) - \ln(1 - x_*)/x_* = 0. \tag{12}$$

  The solution $u(x)$ is determined up to a multiplicative constant.
- On the next step, we maximize in x_* the quantity λ^{-1}. After some simplifications we arrive at the problem

$$F(x_*, x_0) \equiv x_* - 2x_*^2 + x_*^2 x_0^2 \to \max$$

  under the condition (12).
- We show that this problem a single critical point, namely a global maximum.

Statistic $S_{\infty,n}$



Left: $G = 0$ (in yellow) and the Euler–Lagrange equation $\nabla F \parallel \nabla G$ (in blue).

Right: the graph of the maximizer u.

Approximate computation gives $\widehat{x}_* \approx 0.4310514$, $\widehat{x}_0 \approx 0.88889$, and the maximal value equals ≈ 0.20625.

The optimal local index for $S_{\infty,n}$, found in Podkorytova, equals 0.25. Thus, the available local asymptotic efficiency of the test statistics $S_{\infty,n}$ under the shift alternative is 82% from the optimal.