Electronic Journal of Statistics
Vol. 9 (2015) 1738–1743
ISSN: 1935-7524
DOI: 10.1214/15-EJS990

Discussion of "Hypothesis testing by convex optimization"*

Fabienne Comte, Céline Duval and Valentine Genon-Catalot

Université Paris Descartes, MAP5, UMR CNRS 8145

MSC 2010 subject classifications: 62F03.
Keywords and phrases: Convex criterion, hypothesis testing, multiple hypothesis.

Received January 2015.

1. Introduction

We congratulate the authors for this very stimulating paper. Testing statistical composite hypotheses is a very difficult area of the mathematical statistics theory and optimal solutions are found in very seldom cases. It is precisely in this respect that the present paper brings a new insight and a powerful contribution. The optimality of solutions depends strongly on the criterion adopted for measuring the risk of a statistical procedure. In our opinion, the novelty here lies in the introduction of a new criterion different from the usual one (compare criterions (2.1) and (2.2) below). With this new criterion, a minimax optimal solution can be obtained for rather general classes of composite hypotheses and for a vast class of statistical models. This solution is nearly optimal with respect to the usual criterion. The more remarkable results are contained in Theorem 2.1 and Proposition 3.1 and are illustrated by numerous examples.

In what follows, we give some more precise details on the main results necessary to enlighten the strength and the limits of the new theory.

2. The main results

2.1. Theorem 2.1

In this paper, the authors consider a parametric experiment $(\Omega, (P_\mu)_{\mu\in\mathcal{M}})$ where the parameter set $\mathcal{M}$ is a convex open subset of $\mathbb{R}^m$. From one observation ω, it is required to build a test deciding between two composite hypotheses $H_X : \mu \in X$, $H_Y : \mu \in Y$ where X, Y are convex compact subsets of $\mathcal{M}$. Assumptions on the subsets X, Y are thus quite general and the problem is taken as symmetric (no distinction is done between the hypotheses such as choosing H_0 versus H_1). We come back to this point later on.

*Main article 10.1214/15-EJS1054.

A test ϕ is identified with a random variable $\phi : \Omega \to \mathbb{R}$ and the decision rule for H_X versus H_Y is

$$\phi < 0 \Leftrightarrow \text{ reject } H_X, \quad \phi \geq 0 \Leftrightarrow \text{ accept } H_X$$

so that $-\phi$ is a decision rule for H_Y versus H_X.

The theory does not consider *randomized tests* which are found to be optimal tests in the classical theory for discrete observations.

The usual risk of a test is based on the two following errors:

$$R(\phi) := \max\{\epsilon_X(\phi) := \sup_{x \in X} P_x(\phi < 0),\ \epsilon_Y(\phi) := \sup_{y \in Y} P_y(\phi \geq 0)\}. \tag{2.1}$$

We use the notation $\epsilon_X(\phi)$, $\epsilon_Y(\phi)$ to stress on the dependence on ϕ. An optimal minimax test with respect to the classical criterion would be a test achieving

$$\min_{\phi} \max_{(x,y) \in X \times Y} (P_x(\phi < 0) + P_y(\phi \geq 0)).$$

The authors introduce another risk $r(\phi)$, larger than $R(\phi)$ by the simple Markov inequality:

$$r(\phi) := \max\{\sup_{x \in X} \mathbb{E}_x(e^{-\phi}),\ \sup_{y \in Y} \mathbb{E}_y(e^{\phi})\}. \tag{2.2}$$

The main result contained in Theorem 2.1 is that there exists an optimal minimax solution with respect to the augmented criterion: a triplet (ϕ_*, x_*, y_*) exists that achieves

$$\min_{\phi} \max_{(x,y) \in X \times Y} (\mathbb{E}_x(e^{-\phi}) + \mathbb{E}_y(e^{\phi})) := 2 \log \varepsilon_* \tag{2.3}$$

Such a triplet can be explicitly computed in classical examples (discrete, Gaussian, Poisson) together with the value ε_*. Moreover, the usual error probabilities $\epsilon_X(\phi_*), \epsilon_Y(\phi_*)$ can be computed too.

Nevertheless, conditions on the parametric family are required and the optimal test is to be found in a specific class $\mathcal{F}$ of tests. The whole setting (conditions 1.–4.) must be "a good observation scheme" which roughly states:

- First, the parametric family *must be* an exponential family with its natural parameter set: ($'$ denotes the transpose)

$$dP_\mu(\omega) = p_\mu(\omega)\, dP(\omega) = \exp(\mu' S(\omega) - \psi(\mu))\, dP(\omega), \tag{2.4}$$

$$\mathcal{M} = \text{ interior of } \{\mu \in \mathbb{R}^m, \int_\Omega \exp(\mu' S)\, dP < +\infty\}$$

$$S : \omega \in \Omega \to S(\omega) \in \mathbb{R}^m.$$

- The class $\mathcal{F}$ of possible tests is a finite-dimensional linear space of continuous functions, and contains constants and all Neyman-Pearson statistics $\log(p_\mu(.)/p_\nu(.))$. This is not a surprising assumption but means that the class of tests among which an optimal solution is to be found is exactly

$$\mathcal{F} = \{a' S + b, a \in \mathbb{R}^m, b \in \mathbb{R}\} \tag{2.5}$$

- Condition 4. is more puzzling: the class $\mathcal{F}$ must be such that for all $\phi \in \mathcal{F}$,

$$\mu \to \log \int_\Omega e^\phi p_\mu \, dP$$

is defined and concave in $\mathcal{M}$. This means that:

$$\mu \to \log \int_\Omega \exp\left[(a' + \mu')S + b - \psi(\mu)\right] dP$$

is defined and concave in $\mu \in \mathcal{M}$. This may reduce the class $\mathcal{F}$. On the three examples (discrete, Gaussian, Poisson) condition 4. is satisfied. However, on other examples (see the discussion below), condition 4. together with condition 3. requires a comment.

Because of (2.4) and (2.5), the computation of

$$(\phi, x, y) \to \mathbb{E}_x(e^{-\phi}) + \mathbb{E}_y(e^\phi)$$

is easy and the solution of (2.3) is obtained by solving a simple optimization problem. Theorem 2.1 provides the solution $\phi_* = (1/2)\log(p_{x_*}/p_{y_*})$ and a remarkable result is that:

$$\varepsilon_* = \rho(x_*, y_*)$$

where ρ is the Hellinger affinity of P_{x_*}, P_{y_*}.

An important point too concerns the translated detectors $\phi_*^a := \phi_* - a$. Equation (4), p. 4, shows that by a translation, the probabilities of wrong decision can be upper-bounded as follows:

$$\epsilon_X(\phi_*^a) \le e^a \varepsilon_*, \ \epsilon_Y(\phi_*^a) \le e^{-a}\varepsilon_*.$$

Therefore, by an appropriate choice of a, one can easily break the symmetry between hypotheses, choose H_0 and H_1 and reduce one error while increasing the other one.

Examples (2.3.1, 2.3.2, 2.3.3) are particularly illuminating. All computations are easy and explicit.

The extension of the theory to repeated observations is relatively straightforward due to the exponential structure of the parametric model and we will not discuss it.

2.2. Proposition 3.1

Another very powerful result concerns the case where one has to decide not only on a couple of hypotheses but on more than two hypotheses. The testing of unions is an impressive result. The problem is of deciding between:

$$H_X : \mu \in X = \bigcup_{i=1}^m X_i, \ H_Y : \mu \in Y = \bigcup_{i=1}^n Y_i$$

on the basis of one observation $\omega \sim p_\mu$. Here, X_i, Y_j are subsets of the parameter set. The authors consider tests ϕ_{ij} available for the pair (H_{X_i}, H_{Y_j}) such that

$$\mathbb{E}_x(e^{-\phi_{ij}}) \leq \epsilon_{ij}, \forall x \in X,\ \mathbb{E}_y(e^{\phi_{ij}}) \leq \epsilon_{ij}, \forall y \in Y,$$

(for instance the optimal tests of Theorem 2.1, but any other test would suit) and define the matrices $E = (\epsilon_{ij})$ and

$$H = \begin{bmatrix} 0 & E \\ E' & 0 \end{bmatrix}.$$

An explicit test ϕ for deciding between H_X and H_Y is built using an eigenvector of H and the risk $r(\phi)$ of this test is evaluated with accuracy: it is smaller than $||E||_2$ the spectral norm of E.

A very interesting application, which is illustrated on numerous examples, is when $H_X = H_0 : \mu \in X_0$ is one hypothesis (not a union) and $H_Y = \bigcup_{j=1}^n H_j : \mu \in \bigcup_{j=1}^n Y_j$ is a union. One simply builds for $j = 1, \ldots, n$ the optimal tests $\phi_*(H_0, H_j) := \phi_{0j}$ with errors bounded by ϵ_{0j} (obtained, for instance, by Theorem 2.1). The matrix E is the $(1, n)$ matrix $E = [\epsilon_{01}\ \ldots\ \epsilon_{0n}]$. As $E'E$ has rank 1, its only non null eigenvalue is equal to $Tr(E'E) = \sum_{j=1}^n \epsilon_{0j}^2 = (||E||_2)^2$. The eigenvector of H is given by $[1\ h_1\ \ldots\ h_n]'$ with $h_i = \epsilon_{0i}/(\sum_{j=1}^n \epsilon_{0j}^2)^{1/2}$. Then, the optimal test for H_0 versus the union $\bigcup_{j=1}^n H_j$ is explicitly given by

$$\phi = \min_{1 \leq j \leq n} \left\{ \phi_{0j} - \log\left(\epsilon_{0j}/(\sum_{j=1}^n \epsilon_{0j}^2)^{1/2}\right) \right\}.$$

For all $\mu \in X_0$, $\mathbb{E}_\mu(e^{-\phi}) \leq \left(\sum_{j=1}^n \epsilon_{0j}^2\right)^{1/2}$ and for all $\mu \in \bigcup_{j=1}^n Y_j$, $\mathbb{E}_\mu(e^{\phi}) \leq \left(\sum_{j=1}^n \epsilon_{0j}^2\right)^{1/2}$.

Then, given a value ε, if it is possible to tune the test $\phi_{0j} := \phi_{0j}(\varepsilon)$ of H_0 versus H_j to have a risk less than $\varepsilon/\sqrt{n}$, then the resulting test for the union $\phi(\varepsilon)$ has risk less than ε.

This is remarkable: if we consider the test $\min_{1 \leq j \leq n}\{\phi_{0j}\}$ for H_0 versus the union $\bigcup_{j=1}^n H_j$, we have

$$P_0(\min_{1 \leq j \leq n}\{\phi_{0j}\}) \leq \sum_{j=1}^n \epsilon_{0j}.$$

To get a risk bounded by ε, one would have to tune the test ϕ_{0j} of H_0 versus H_j to have a risk less than ε/n.

3. Discussion

- The theory is restricted to exponential families of distributions with natural parameter space. As noted by the authors in the introduction, this is the price to pay for having very general hypotheses. The problem of finding more general statistical experiments which would fit in the theory

is open and worth investigating. The new risk criterion seems to be a more flexible one for finding new optimal solutions.

- The fact that the sets X, Y must be compact sets is restrictive. We wonder if there are possibilities to weaken this constraint.
- The combination of conditions 3. and 4. on the class of tests implies a reduction of the class of tests. Consider the case of exponential distributions, where

$$p_\mu(\omega) = \mu \exp(-\mu\,\omega), \quad \omega \in (0, +\infty), \quad \mu \in \mathcal{M} = (0, +\infty), \quad S(\omega) = -\omega.$$

By condition 3., $\mathcal{F}$ contains $\log p_\mu/p_\nu$ for all $\mu, \nu > 0$, hence $\mathcal{F}$ contains all tests $\phi(\omega) = a\omega + b, a, b \in \mathbb{R}$.
For condition 4., to compute $F(\mu)$ the condition $\mu > a$ is required and

$$F(\mu) = \log\Big[\int_0^{+\infty} \mu \exp((a-\mu)\omega + b) d\omega\Big] = b + \log[\mu/(\mu - a)].$$

As $F''(\mu) = a(2\mu - a)/(\mu^2(\mu - a)^2)$, F is concave if and only if $a \leq 0$. Therefore, condition 4. restricts the class of tests to $\mathcal{F} = \{\phi = a\omega + b, a \leq 0, b \in \mathbb{R}\}$. This raises a contradiction: condition 3) states that $\mathcal{F}$ must contain all $\log p_\mu/p_\nu$, thus all $\phi = a\omega + b, a, b \in \mathbb{R}$. We wonder if condition 4. could be stated differently so as to avoid this contradiction.

- Generally, when testing statistical hypotheses, one chooses a hypothesis H_0 and builds a test of H_0 versus H_1. One wishes to control the error of rejecting H_0 when it is true and the other error does not really matter. A discussion on this point is lacking in relation with Theorem 2.1, formula (4).
- Randomized tests are not considered here. However, they are found as optimal solutions in the fundamental Neyman-Pearson lemma. In estimation theory, randomized estimators are of no use because, due to the convexity of loss functions, a non randomized estimator is better. In the setting of the paper, is there such a reason to eliminate randomized tests?
- The criterion $r(\phi)$ is larger than the usual one, entailing a loss. The notion of "provably optimal test" introduced in this study is not commented in the text. So, it is difficult to understand or quantify it. Maybe more comments on Theorem 2.1 (ii) would help. At this point, let us notice that the notation ϵ_X without dependence on the test statistic ϕ is a bit misleading. It should be $\epsilon_X(\phi)$, so that in formula (4) of Theorem 2.1, we would read $\epsilon_X(\phi_*)$.
Another point is the comparison between $\epsilon_X(\phi_*)$ and ε_*. Apart from the Markov inequality, would it be possible to quantify $\varepsilon_* - \epsilon_X(\phi_*)$? Example 2.3.1 give both quantities without comment. Examples 2.3.2 and 2.3.3 only give ε_*.
- We looked especially at Section 4.2. The Poisson case is illuminating and helps understanding the application of Proposition 3.1. On the contrary, we had difficulties with Section 4.2.3 (Gaussian case) and do not understand why the optimal test of H_0 versus H_j is not computed as in Section

2.3.1. Consequently, more details on the proof of Proposition 4.2 would have been useful. Also, in these examples, are the quantities ϵ_{0j} computed as $\varepsilon_*(\phi_*(H_0, H_j))$ or as $\epsilon_{X_0}(\phi_*(H_0, H_j))$? The numerical illustrations were also difficult to follow.

4. Concluding remarks

To conclude, the criterion $r(\phi)$, defined in (2.2), provides a powerful new tool to build nearly optimal tests for multiple hypotheses. The large number of concrete and convincing examples is impressive. We believe that this paper offers a new insight on the theory of testing statistical hypotheses and will surely inspire new research and new developments.

References

[1] Goldenshluger, A., Juditski, A., and Nemirovski, A. (2014). Hypothesis testing by convex optimization. *Arxiv preprint 1311.6765v6*.