Electronic Journal of Statistics
Vol. 9 (2015) 1713–1719
ISSN: 1935-7524
DOI: 10.1214/15-EJS1014

Discussion of "Hypotheses testing by convex optimization"*

Lucien Birgé

Sorbonne Universités – UPMC Université Paris 06 – U.M.R. 7599, L.P.M.A.
Case courrier 188, Université Paris 06, 4 Place Jussieu, F-75252 Paris Cedex 05, France,
e-mail: lucien.birge@upmc.fr

Received March 2015.

I was very happy to read this paper by Sasha, Anatoli and Arkadi not only because it is an exciting paper about testing problems that I have been interested in for many years but also for some more personal reasons. It actually reminded me of various problems about testing between sets that I begun to consider and work on almost fourty years ago and to which I gave a continuous interest up to now. It also brought back to my memory many exchanges and discussions that I had in the late seventies and early eighties with Lucien Le Cam, as well as many seminars and talks about robustness, which was at that time a very fashionable subject, and many friends that I made then like Tadeusz Bednarski, Gabor Tusnády and Piet Groeneboom, among others.

1. Historical remarks

The general problem of testing between two non-trivial sets received a lot of attention in the seventies along two different streams of research. One was initiated by Le Cam and its collaborators much earlier, actually in the fifties and a milestone paper was Kraft (1955) about the consistency of tests. It actually contains (Theorem 5) a fundamental result by Le Cam (previously unpublished) that I shall comment about below and which provides the performance of a best test between two convex sets of probabilities. Other important results about the performance of tests are provided by Le Cam (1973) in a paper which was directed towards the use of tests in order to derive "universal" estimators under some dimensionality restrictions.

Le Cam's work was about the possibility of testing efficiently between two sets of probabilities with application to estimation while the theory of robustness was about a different problem that I could summarize by "improving the stability" of statistical procedures. When applied to testing between two simple hypotheses, say $\{P\}$ and $\{Q\}$, it amounts to finding tests that are more or less equivalent to the classical likelihood ratio (Neyman-Pearson) tests between P and Q but with errors that do not increase too much when the truth is actually slightly different

*Main article 10.1214/15-EJS1054.

from either P or Q. This amounts to finding tests between some vicinities $\mathcal{P}$ and $\mathcal{Q}$ of P and Q respectively, the result depending on the topology that is used. Various results in this direction appeared in the sixties and seventies after the milestone paper by Huber (1965). It would be much too long to cite them all but Huber showed how to find explicit optimal tests between $\mathbb{L}_1$-balls, among other vicinities, and more generally between sets that are dominated by two-alternating capacities – Huber and Strassen (1973). There are indeed many available results about tests between convex sets but a large part of them is of a purely theoretical nature (existence results) and does not provide explicit tests that perform as predicted by the theory. It is a great merit of this paper to provide such tests in some interesting and useful statistical frameworks.

2. Kraft and Le Cam's results

Let us begin with some elementary facts and notations. To test with a random variable X between two probability sets $\mathcal{P}$ and $\mathcal{Q}$, we use a test function φ with values in $\{-1,1\}$, deciding $\mathcal{P}$ when $\varphi(X)=1$ and $\mathcal{Q}$ otherwise. This results in errors of the form

$$\alpha(\mathcal{P},\varphi)=\sup_{P\in\mathcal{P}}\mathbb{P}_P[\varphi(X)=-1] \qquad \text{and} \qquad \alpha(\mathcal{Q},\varphi)=\sup_{Q\in\mathcal{Q}}\mathbb{P}_Q[\varphi(X)=1].$$

For a given test φ these errors do not change if we replace both $\mathcal{P}$ and $\mathcal{Q}$ by their convex hulls, hereafter denoted by $\mathrm{Co}(\mathcal{P})$ and $\mathrm{Co}(\mathcal{Q})$ respectively. Le Cam was interested by what he called the "testing affinity"

$$\pi(\mathcal{P},\mathcal{Q})=\inf_{\varphi}\{\alpha(\mathcal{P},\varphi)+\alpha(\mathcal{Q},\varphi)\}=\pi\big(\mathrm{Co}(\mathcal{P}),\mathrm{Co}(\mathcal{Q})\big)\,,$$

(where the infimum runs over all test functions φ) which measures, in a sense, the performance of a best test between $\mathcal{P}$ and $\mathcal{Q}$. In particular, denoting by dP and dQ the densities of P and Q with respect to any dominating measure,

$$\pi(P,Q)=\pi(\{P\},\{Q\})=\int\min\{dP,dQ\}=1-D(P,Q),$$

where $D(P,Q)$ denotes the "variation distance":

$$D(P,Q)=\sup_A|P(A)-Q(A)|=\frac{1}{2}\int|dP-dQ|.$$

The fundamental theoretical result of Le Cam which appeared in Kraft (1955) says:

Theorem 1 *If $\mathcal{P}$ and $\mathcal{Q}$ are two sets of probabilities, then*

$$\begin{aligned}\pi(\mathcal{P},\mathcal{Q}) &= 1-D\big(\mathrm{Co}(\mathcal{P}),\mathrm{Co}(\mathcal{Q})\big) = 1-\inf_{P\in\mathrm{Co}(\mathcal{P}),\,Q\in\mathrm{Co}(\mathcal{Q})}D(P,Q)\\ &= \sup_{P\in\mathrm{Co}(\mathcal{P}),\,Q\in\mathrm{Co}(\mathcal{Q})}\pi(P,Q). \qquad (2.1)\end{aligned}$$

In words, the testing affinity between two convex sets is determined by their variation distance.

Although quite precise, this result is actually very difficult to use for analyzing concrete testing problems for two reasons. First the proof of Theorem 1 is based on the Hahn-Banach Theorem and therefore does not provide the construction of an optimal test. Then many problems involve i.i.d. observations $X_1, \ldots, X_n$ with joint distribution $R^{\otimes n}$ with R belonging either to $\mathcal{P}$ or $\mathcal{Q}$ so that the application of Theorem 1 requires to compute $\pi(S,T)$ for S and T in the convex hulls of $\mathcal{P}^{\otimes n} = \{P^{\otimes n}, P \in \mathcal{P}\}$ and $\mathcal{Q}^{\otimes n} = \{Q^{\otimes n}, Q \in \mathcal{Q}\}$. Unfortunately, there is no direct relationship between $\pi(P^{\otimes n}, Q^{\otimes n})$ and $\pi(P,Q)$.

A major idea of Kraft and Le Cam, in view of solving the second problem, was the introduction of the Hellinger distance h and affinity ρ as substitutes to π and D:

$$\rho(P,Q) = \int \sqrt{dP dQ} = 1 - h^2(P,Q) = 1 - \frac{1}{2}\int \left(\sqrt{dP} - \sqrt{dQ}\right)^2.$$

They can be related to π or D via the following inequalities:

$$\pi(P,Q) \le \rho(P,Q) \le \sqrt{\pi(P,Q)[2 - \pi(P,Q)]}, \tag{2.2}$$

or equivalently,

$$D(P,Q) \ge h^2(P,Q) \ge 1 - \sqrt{1 - D^2(P,Q)]} \ge (1/2)D^2(P,Q).$$

The main advantage of ρ over π derives from the fact that $\rho(P^{\otimes n}, Q^{\otimes n}) = \rho^n(P,Q)$ which allows to deal with i.i.d. samples. Let us now define, for two probability sets $\mathcal{P}$ or $\mathcal{Q}$,

$$\rho(\mathcal{P},\mathcal{Q}) = \sup_{P \in \mathrm{Co}(\mathcal{P}),\, Q \in \mathrm{Co}(\mathcal{Q})} \rho(P,Q) = 1 - \inf_{P \in \mathrm{Co}(\mathcal{P}),\, Q \in \mathrm{Co}(\mathcal{Q})} h^2(P,Q).$$

It then follows from (2.1) and (2.2) that

$$\pi(\mathcal{P},\mathcal{Q}) \le \rho(\mathcal{P},\mathcal{Q}) \qquad \text{and} \qquad \pi\left(\mathcal{P}^{\otimes n}, \mathcal{Q}^{\otimes n}\right) \le \rho\left(\mathcal{P}^{\otimes n}, \mathcal{Q}^{\otimes n}\right).$$

Moreover, the following fundamental result holds for ρ.

Theorem 2 *If $\mathcal{P}$ and $\mathcal{Q}$ are two sets of probabilities, then*

$$\rho\left(\mathcal{P}^{\otimes n}, \mathcal{Q}^{\otimes n}\right) \le \rho^n(\mathcal{P},\mathcal{Q}) = \sup_{P \in \mathrm{Co}(\mathcal{P}),\, Q \in \mathrm{Co}(\mathcal{Q})} \rho^n(P,Q).$$

Putting everything together, we can conclude that if $\mathcal{P}$ and $\mathcal{Q}$ are convex, then

$$\begin{aligned} \pi\left(\mathcal{P}^{\otimes n}, \mathcal{Q}^{\otimes n}\right) &\le \rho^n(\mathcal{P},\mathcal{Q}) = \sup_{P \in \mathcal{P},\, Q \in \mathcal{Q}} \rho^n(P,Q) \\ &= \left[1 - h^2(\mathcal{P},\mathcal{Q})\right]^n \le \exp\left[-n h^2(\mathcal{P},\mathcal{Q})\right], \end{aligned}$$

with $h^2(\mathcal{P},\mathcal{Q}) = \inf_{P \in \mathcal{P},\, Q \in \mathcal{Q}} h^2(P,Q)$. This shows that, as soon as two convex sets of probabilities $\mathcal{P}$ and $\mathcal{Q}$ are separated (i.e. $h(\mathcal{P},\mathcal{Q}) > 0$), the errors of a best test between $\mathcal{P}$ and $\mathcal{Q}$ decrease exponentially fast with the number n of observations. But this does not solve the problem of finding explicit tests with such a performance.

3. An alternative point of view

The point of view that underlines the paper by Sasha, Anatoli and Arkadi derives from the following observations. The best test between P and Q is the likelihood ratio test given by $\varphi(x) = -1$ if $L(x) = \log(dQ/dP)(x) > 0$ and $\varphi(x) = 1$ if $L(x) = \log(dQ/dP)(x) < 0$, the value of $\varphi(x)$ when $L(x) = 0$ being irrelevant. Hence

$$\pi(P,Q) = \mathbb{P}_P[\varphi(X) = -1] + \mathbb{P}_Q[\varphi(X) = 1] \le \rho(P,Q).$$

But it is actually much more fruitful to proceed differently in order to bound both errors of this test separately. The following sequence of inequalities is straightforward but nevertheless enlightening.

$$\begin{aligned}\mathbb{P}_P[\varphi(X) = -1] &\le \mathbb{P}_P[L(X) \ge 0] \le \inf_{\lambda>0} \mathbb{E}_P\left[e^{\lambda L(X)}\right] \\ &\le \mathbb{E}_P\left[e^{L(X)/2}\right] = \mathbb{E}_P\left[\sqrt{(dQ/dP)(X)}\right] = \rho(P,Q)\end{aligned}$$

and also

$$\mathbb{P}_Q[\varphi(X) = 1] \le \mathbb{P}_Q[L(X) \le 0] \le \mathbb{E}_Q\left[e^{-L(X)/2}\right] = \rho(P,Q).$$

This actually leads to the suboptimal bound $\pi(P,Q) \le 2\rho(P,Q)$ but this is unimportant in the very interesting case of a small value of $\rho(P,Q)$. Here we have an exemple of a function $\psi(x) = \exp[L(x)/2]$ such that $\mathbb{E}_P[\psi(X)] \le \rho(P,Q)$ and $\mathbb{E}_Q[1/\psi(X)] \le \rho(P,Q)$. The authors actually call $\phi = \log\psi$ a *detector*. Inequalities of the form

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[\psi(X)] \le \alpha < 1 \qquad \text{and} \qquad \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[1/\psi(X)] \le \beta < 1 \tag{3.1}$$

are indeed extremely useful to control the performance of tests between $\mathcal{P}$ and $\mathcal{Q}$ based on the detector ϕ according to the following (trivial) lemma which already appears (hidden in the proofs) in Birgé (1984), but also in earlier works like Chernoff (1952) about large deviations.

Lemma 1 *Let $\mathcal{P}_i$ and $\mathcal{Q}_i$, $1 \le i \le n$ be sets of probabilities on measurable spaces $\mathcal{X}_i$ and $\mathcal{P} = \{\bigotimes_{i=1}^n P_i, P_i \in \mathcal{P}_i\}$, $\mathcal{Q} = \{\bigotimes_{i=1}^n Q_i, Q_i \in \mathcal{Q}_i\}$ the corresponding sets of probabilities on $\prod_{i=1}^n \mathcal{X}_i$. If there exists for each i a function ψ_i such that*

$$\sup_{P\in\mathcal{P}_i} \mathbb{E}_P[\psi_i(X)] \le \alpha_i \le 1 \qquad \textit{and} \qquad \sup_{Q\in\mathcal{Q}_i} \mathbb{E}_Q[1/\psi_i(X)] \le \beta_i \le 1, \tag{3.2}$$

then, for all $y \in \mathbb{R}$ and random variables $X_i \in \mathcal{X}_i$ for $1 \le i \le n$,

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\left[\sum_{i=1}^n \log\psi_i(X_i) \ge y\right] \le \exp\left[-y + \sum_{i=1}^n \log\alpha_i\right]$$

and

$$\sup_{Q\in\mathcal{Q}} \mathbb{P}_Q\left[\sum_{i=1}^n \log\psi_i(X_i) \le y\right] \le \exp\left[y + \sum_{i=1}^n \log\beta_i\right].$$

It follows from this lemma that, once one has found a set of functions ψ_i, one can easily derive tests between the convex envelopes of $\mathcal{P}$ and $\mathcal{Q}$ with controlled

errors and play with y in order to balance between them. Finding suitable functions ψ_i is therefore essential. This is the approach that I considered in Birgé (1984) and more recently in Birgé (2013). But again, given two sets $\mathcal{P}$ and $\mathcal{Q}$, the problem of finding ψ satisfying (3.1) has no explicit solution in general, in contrast to Huber's results that provide explicit tests between some particular vicinities of P and Q. Unfortunately his results do not apply to Hellinger vicinities which are not generated by two-alternating capacities but, fortunately, although abstract in the general case, my results apply to Hellinger balls and provide in this particular case some explicit tests which are actually likelihood ratio tests between the closest points in the balls. This is exactly the type of result that Sasha, Anatoli and Arkadi get for their "good observation schemes". One can readily see from their illustrations (a), (b) and (c) that the "least favorable" pair (P_{x_*}, P_{y_*}) is actually a pair that minimizes the Hellinger distance between the two hypotheses and the "nearly optimal" test is a likelihood ratio test between them.

As we have seen, finding pairs of sets $\mathcal{P}$ and $\mathcal{Q}$ for which one can explicitly compute a function ψ that satisfies (3.1) is a major issue in testing theory. The interest and importance of the present paper is that it solves this problem for particular sets $\mathcal{P}$ and $\mathcal{Q}$ connected to some classical parametric statistical models. On the one hand the results only apply to some very specific cases (the good observation schemes), on the other hand they allow to derive concrete and practical tests with excellent performances, which is definitely more useful than a mere existence theorem. As we often say in French "on ne peut avoir le beurre et l'argent du beurre !" (no free lunch!). Moreover, in view of Lemma 1, these tests apply to i.i.d. samples and allow to balance between the two kinds of errors. Although dealing with some particular parametric models, they nevertheless apply to many interesting situations as illustrated by the authors in their Section 4.

4. Combining elementary tests

An important part of the paper (namely Section 3) is devoted to various ways of combining detectors $\phi = \log \psi$, where ψ satisfies (3.1), in order to test between more complex hypotheses. Given a family $H_1, \dots, H_M$ of hypotheses (with H_j corresponding to $P \in \mathcal{P}_j$) for which one can find tests between pairs (H_i, H_j), $1 \le i < j \le M$, it is not obvious to design *good* tests between $\cup_{j=1}^{m} H_i$ and $\cup_{j=m+1}^{M} H_i$. I am personally not fond of considering this problem when some of the hypotheses overlap or, more generally, when some of them are almost indistinguishable. Le Cam (1973) (and this also follows from the inequalities of the previous Section 2) shows that if $nh^2(\mathcal{P}, \mathcal{Q})$ is too small, no test can correctly distinguish between $\mathcal{P}^{\otimes n}$ and $\mathcal{Q}^{\otimes n}$. I therefore believe that when combining tests between various hypotheses H_j, one should not try to test between H_i and H_j when these hypotheses are too close (in Hellinger distance). In such a case, one knows that there is no hope to get small errors so that such a situation should be avoided and a good solution is to proceed as indicated in Section 3.2.1 and avoid testing between hypotheses that are too close.

Considering the problem of choosing between M hypotheses (but this remark is also valid for testing, when $M = 2$) I believe that it has to be put in the classical decision-theoretic framework with a decision function $\delta(X)$ with values in $\{1, \ldots, M\}$. Since the assumption is that the true distribution satisfies at least one assumption the decision δ should choose one. This rules out any strategy, like some that appear in Section 3 which, may be, reject all hypotheses and therefore lead to no decision. A simple solution, which would not increase the rejection rate and therefore the errors, would be, for instance, to decide at random when the multiple testing procedure rejects all hypotheses.

Clearly, the procedure of Section 3.2.1 tends to improve the situation. It can be viewed as the choice of a family of "pseudo"-distances between the different hypotheses, the "distance" between H_i and H_j being zero when $(i, j) \in \mathcal{C}$ (assuming symmetry: $(i, j) \in \mathcal{C}$ is equivalent to $(j, i) \in \mathcal{C}$) and one otherwise. For a given i, one only tests with H_j if the "distance" between H_i and H_j is one. One could actually adopt a more sophisticated strategy with mutual distances between the assumptions being arbitrary nonnegative numbers. One could for instance take, as the "distance" between H_i and H_j, minus the logarithm of the error of a best test between them. To build the final decision function, one should not only consider the various tests involved but also look at the mutual distances between the hypotheses and decide according to these mixed informations. The idea is to decide for an $\widehat{i}$ such that no test for which H_j is far from $H_{\widehat{i}}$ rejects it, based on the fact that two hypotheses that are close cannot be properly distinguished while two hypotheses that are far apart lead to a test with small errors. I actually used this argument in Birgé (1983) and Birgé (2006) to derive estimators from families of tests. I am convinced that the method that I used to build T-estimators can be adapted to deal with multiple hypotheses via the function $\mathcal{D}_{\mathbf{X}}$ provided by (4.5) of Birgé (2006). One could analogously use a suitable version of $\mathcal{D}_{\mathbf{X}}(k)$ as a criterion of the "credibility" of the assumption H_k (the larger $\mathcal{D}_{\mathbf{X}}(k)$ the less credible H_k) and finally decide for the minimizer over k of $\mathcal{D}_{\mathbf{X}}(k)$. Various modifications of this procedure are certainly possible.

5. Conclusion

I definitely find this paper exciting and hope it will open a new research trend towards finding explicit detectors for testing between two hypotheses for other observation schemes. I hope that more examples will be found in the future. I also greatly appreciated the applications of Section 4 but discussing this aspect is not really in my field of expertise. As to the problem of handling many hypotheses, I believe that the point of view developed in Section 3.2.1 is the more fruitful and promising one and that the authors should pursue in this direction.

References

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237. MR0722129

Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3(2):259–282. MR0764150

Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325. MR2219712

Birgé, L. (2013). Robust tests for model selection. In Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V., and Maathuis, M. H., editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes*, volume 9, pages 47–64. IMS Collections. MR3186748

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507. MR0057518

Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.*, 36:1753–1758. MR0185747

Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.*, 1:251–263. MR0356306

Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.*, 2:125–141. MR0073896

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53. MR0334381