A note on robust hypothesis testing

Luc Devroye * László Györfi † Gábor Lugosi ‡

July 27, 2001

Abstract

We introduce a simple new hypothesis testing procedure, which, based on an independent sample drawn from a certain density, detects which of k nominal densities is the true density is closest to, under the total variation (L_1) distance. We obtain a density-free uniform exponential bound for the probability of false detection.

KEY WORDS AND PHRASES: robust hypotheses testing, robust detection

*School of Computer Sciences, McGill University, Montreal, Canada H3A 2K6. e-mail: luc@kriek.cs.mcgill.ca

†Dept. of Computer Science and Information Theory, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary. e-mail: gyorfi@inf.bme.hu

‡Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. e-mail: lugosi@upf.es. The research of this author was supported by DGI grant BMF2000-0807

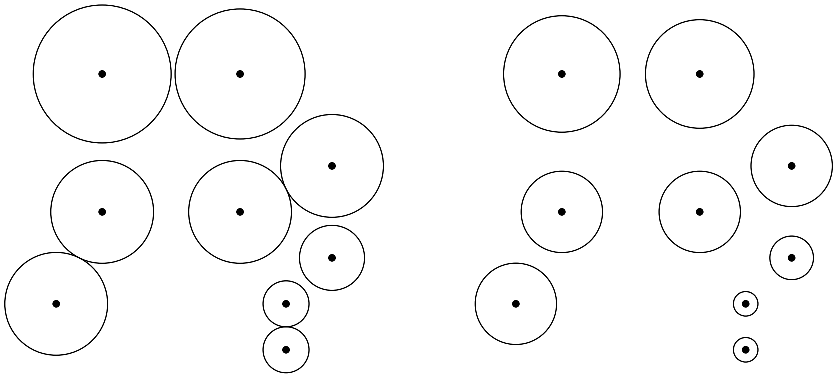

Figure 1: The hypothesis classes H_j are illustrated here for $k = 9$ with $\epsilon = 0$ on the left and $\epsilon > 0$ on the right. The centers of the balls represent the nominal densities $f^{(j)}$.

1 Result

A model of robust hypothesis testing may be formulated as follows: let $f^{(1)}, \dots, f^{(k)}$ be fixed densities on $\mathbb{R}^d$ which are the nominal densities under k hypotheses. We observe i.i.d. random vectors $X_1, \dots, X_n$ according to a common density f. Under the hypothesis H_j ($j = 1, \dots, k$) the density f is a distorted version of $f^{(j)}$. This notion may be formalized in various ways. In this note we assume that the true density lies within a certain total variation distance of the underlying nominal density. More precisely, we assume that there exists a positive number ϵ such that for some $j \in \{1, \dots, k\}$

$$||f - f^{(j)}|| \leq \Delta_j - \epsilon,$$

where $\Delta_j \stackrel{\text{def}}{=} (1/2) \min_{i \neq j} ||f^{(i)} - f^{(j)}||$. Here $||f - g|| = \int |f - g|$ denotes the L_1 distance between two densities. Thus, we formally define the k hypotheses by

$$H_j = \left\{ f : ||f - f^{(j)}|| \leq \Delta_j - \epsilon \right\}, \quad j = 1, \dots, k .$$

The goal is to construct tests which, with high probability, assign to the observed sample the index j of the correct nominal density, that is, determines to which hypothesis H_j the density f belongs to.

Perhaps the most standard testing method is maximum likelihood, which accepts the j-th nominal density $f^{(j)}$ if

$$\prod_{\ell=1}^{n} \frac{f^{(j)}(X_\ell)}{f^{(i)}(X_\ell)} > 1 \quad \text{for all } i \neq j \ .$$

It is easy to see that this method is not robust in the sense that arbitrarily small deviations from the nominal density may cause a catastrophic behavior. We provide a simple example in Section 2 below. In the special case when $k = 2$, a remarkable result of Huber [11] shows that a simple modification of the maximum likelihood test is optimal in the minimax sense, that is, it minimizes the worst-case probability of error in the given model. More precisely, Huber's test uses the modified likelihood ratio

$$\prod_{\ell=1}^{n} \max\left[c, \min\left(c', \frac{f^{(1)}(X_\ell)}{f^{(2)}(X_\ell)}\right)\right]$$

for constants c, c' which depend on the nominal densities. The disadvantage of Huber's test is that the values of these constants are given implicitely only and determining them may be problematic, especially when $d > 1$. Also, Huber's result does not cover the case $k > 2$ and it does not provide nonasymptotic bounds for the probability of error.

Other attempts for constructing robust tests involve nonparametric estimates of the underlying density f and deciding based on its distance from the nominal densities (see, e.g., Györfi and van der Meulen [8], Zabin and Wright [14]). However, due to the fact that the L_1 error of any density estimate is bounded away from zero for some densities at any sample size (see [2]), it seems unlikely that the error of these tests can be bounded uniformly for any $f \in \cup_{j=1}^{k} H_j$, or at least such a result seems to be very difficult to prove.

The purpose of this paper is to introduce a new, simple, explicit testing procedure with a uniform nonasymptotic exponential bound for the probability of error. For surveys on robust statistics we refer the reader to Huber [11] and Hampel et al. [9].

In order to define the proposed test, introduce the empirical measure

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \in A} \ ,$$

where $\mathbb{I}$ denotes the indicator function and A is a Borel set. Let $\mathcal{A}$ denote the collection of $k(k-1)/2$ sets of the form

$$A_{i,j} = \left\{x : f^{(i)}(x) > f^{(j)}(x)\right\} \ , \quad 1 \le i < j \le k \ .$$

The proposed test is the following: accept hypothesis H_j if

$$\max_{A\in\mathcal{A}} \left| \int_A f^{(j)} - \mu_n(A) \right| = \min_{i=1,\ldots,k} \max_{A\in\mathcal{A}} \left| \int_A f^{(i)} - \mu_n(A) \right| \ .$$

(In case there are several indices achieving the minimum, choose the smallest one.) The main result of this note is the following:

Theorem 1 *For any* $f \in \bigcup_{j=1}^{k} H_j$

$$\mathbb{P}\{\mathit{error}\} \le 2k(k-1)^2 e^{-n\epsilon^2/2}.$$

Proof. Without loss of generality, assume that $f \in H_1$. Recall that by Scheffé's theorem half of the L_1 distance equals the total variation distance:

$$\|f-g\| = 2 \sup_{A\subset\mathbb{R}^d} \left| \int_A f - \int_A g \right| = 2\left(\int_{\{x:f(x)>g(x)\}} f - \int_{\{x:f(x)>g(x)\}} g \right) \ ,$$

where the supremum is taken over all Borel sets of $\mathbb{R}^d$. By Scheffé's theorem,

$$\begin{aligned}
2\max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| &\le \|f - f^{(1)}\| \\
&\le \Delta_1 - \epsilon \\
&\le \frac{1}{2}\|f^{(1)} - f^{(j)}\| - \epsilon \\
&= \max_{A\in\mathcal{A}} \left| \int_A f^{(1)} - \int_A f^{(j)} \right| - \epsilon \\
&\le \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| + \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(j)} \right| - \epsilon
\end{aligned}$$

by the triangle inequality. Rearranging the obtained inequality, we get that

$$\max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| \le \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(j)} \right| - \epsilon \ .$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\{\text{error}\} &= \mathbb{P}\left\{\exists j > 1 : \max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right|\right\} \\
&\le (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right|\right\} \\
&= (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right. \\
&\qquad\qquad \left. < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right\} \\
&\le (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(j)}\right| + \epsilon\right. \\
&\qquad\qquad \left. < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right\} \\
&\qquad \text{(by the inequality derived above)} \\
&\le (k-1)\max_{j>1}\mathbb{P}\left\{\left|\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(j)}\right|\right| > \frac{\epsilon}{2}\right\} \\
&\quad +(k-1)\mathbb{P}\left\{\left|\max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right| > \frac{\epsilon}{2}\right\} \\
&\le 2(k-1)\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f - \mu_n(A)\right| > \frac{\epsilon}{2}\right\} \\
&\qquad \text{(by a double application of the triangle inequality)} \\
&\le 2(k-1)|\mathcal{A}|\max_{A\in\mathcal{A}}\mathbb{P}\left\{\left|\int_A f - \mu_n(A)\right| > \frac{\epsilon}{2}\right\} \\
&\le 2k(k-1)^2 e^{-n\epsilon^2/2} ,
\end{aligned}
$$

where in the last step we used Hoeffding's inequality [10]. □

2 Discussion

METHODOLOGY. The methodology of the proposed test is close in spirit to Yatracos' minimum distance parametric density estimate, see Yatracos [13], Devroye and Lugosi [5, 6, 7].

COMPUTATION. The hypothesis testing method proposed above is computationally quite simple. The sets $A_{i,j}$ and the integrals $\int_A f^{(j)}$ may be computed and stored before seeing the data. Then one merely needs to calculate $\mu_n(A)$ for all $A \in \mathcal{A}$ and compute the test statistics requiring $O(nk^2 + k^2 \log k)$ time. In many applications $k = 2$. In these cases the test becomes especially simple as the class $\mathcal{A}$ contains just one set.

ROBUSTNESS. Note that the theorem does not require any assumption for the nominal densities. (In fact, the result may be formulated in a similar fashion without even assuming the existence of the densities.) The test is robust in a very strong sense: we obtain uniform exponential bounds for the probability of failure under the sole assumption that the distorted density remains within a certain total variation distance of the nominal density.

ADDITIVE NOISE. We illustrate the power of the proposed method on a very simple example showing that the test has an exponentially small probability of error if the nominal density is corrupted by an arbitrary additive noise of a sufficiently small support. Consider k nominal densities $f^{(1)}, \ldots, f^{(k)}$ and assume that the observations are distributed according to one of the nominal densities corrupted by an additive noise. Thus, assume that the X_i's are distributed according to density $f = f^{(1)} \star g$, where the nominal density $f^{(1)}$ is now assumed to be Lipschitz (i.e., $|f^{(1)}(x) - f^{(1)}(y)| \leq c|x - y|$ for some $c > 0$ for all $x, y \in \mathbb{R}$), supported on the bounded set $[-M, M]$, and the density g of the additive noise is assumed to have support in the interval $[-r, r]$, where r is thought of as a small number. The other $k - 1$ nominal densities are arbitrary. Then, according to the theorem, the proposed test is correct with probability larger than $1 - 2k(k-1)^2 e^{-n\epsilon^2/2}$ as long as $\|f - f^{(1)}\| \leq \Delta_1 - \epsilon$. But

$$
\begin{aligned}
\|f - f^{(1)}\| &= \int \left| \int f^{(1)}(x-y) g(y)\,dy - \int f^{(1)}(x) g(y)\,dy \right| dx \\
&\leq \int \int |f^{(1)}(x-y) - f^{(1)}(x)| g(y)\,dy\,dx
\end{aligned}
$$

$$\begin{aligned} &\le \int_{-M-r}^{M+r} \int c|y|g(y)\,dy\,dx \\ &\le 2c(M+r)r\ . \end{aligned}$$

Thus, the condition is satisfied if r is so small that $r \le (\Delta_1 - \epsilon)/2c(M+r)$. This is the only assumption on the noise density g, otherwise it may be completely arbitrary! (Note that boundedness of the support of g is not a necessary condition; we assumed it to simplify the example.)

MAXIMUM LIKELIHOOD DOES NOT WORK. Here we show a simple example to demonstrate that the maximum likelihood test does not share the proved property of the proposed test. Indeed, consider the case when $k = 2$, and the two nominal densities are standard normal and standard Cauchy densities, that is,

$$f^{(1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{and} \quad f^{(2)}(x) = \frac{1}{\pi(1+x^2)}\ .$$

Assume that the data are distributed according to $f = f^{(1)} \star g_c$, where the density of the additive noise is Cauchy:

$$g_c(x) = \frac{1}{\pi c(1+(x/c)^2)}\ ,$$

where c is a small positive constant. It is well known (see, e.g., [4]) that $||f^{(1)} - f^{(1)} \star g_c|| \to 0$ as $c \to 0$, and therefore, for a sufficiently small c, the L_1 distance between f and $f^{(1)}$ can be made arbitrarily small, in particular, $||f^{(1)} - f|| < ||f^{(1)} - f^{(2)}||/2 - \epsilon$. Nevertheless, it is easy to show that for any small c, the probability of error of the maximum likelihood detector converges to one. Indeed, on the one hand,

$$\begin{aligned} \mathbb{E}\left\{ \frac{1}{n} \sum_{\ell=1}^{n} \log f^{(1)}(X_\ell) \right\} &= \int f(x) \log f^{(1)}(x)\,dx \\ &= -\log\sqrt{2\pi} - \frac{1}{2}\int f(x)x^2\,dx \\ &= -\infty\ , \end{aligned}$$

and on the other hand,

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{n}\sum_{\ell=1}^{n}\log f^{(2)}(X_\ell)\right\} &= \int f(x)\log f^{(2)}(x)\,dx \\
&= -\log\pi - \int f(x)\log(1+x^2)\,dx \\
&> -\infty\ .
\end{aligned}$$

Therefore, the strong law of large numbers implies that for sufficiently large n, the maximum likelihood detector errs with probability one.

TESTS BASED ON DENSITY ESTIMATES. An alternative way of performing robust tests is based on estimating the density. Indeed, such methods have been proposed in the literature, see Györfi and van der Meulen [8], Zabin and Wright [14]. These tests cannot compete with the simplicity of the proposed method, and no uniform exponential bound for their probability of error is available. However, hypothesis testing based on density estimates may be necessary if even larger hypothesis classes need to be considered. A stronger notion of robust hypothesis testing is obtained if one requires good testing whenever the true density is closer to the nominal density than to any other density in the finite collection. Formally, this leads to the hypotheses

$$\overline{H}_j = \{f : \|f - f^{(j)}\| < \min_{i\neq j}\|f - f^{(i)}\|\}, \quad j = 1, \ldots, k\ ,$$

that is, the sets $\overline{H}_j$ form a Voronoi partition of the set of all densities. This problem may be solved by using a nonparametric estimate f_n of f and accepting $\overline{H}_j$ if $\|f_n - f^{(j)}\|$ is minimal among the $\|f_n - f^{(i)}\|$, $i = 1, \ldots, k$. (Break ties by selecting the smallest index.) A suitable choice is the kernel estimate defined by

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where $K : \mathbb{R}^d \to \mathbb{R}^+$ is a fixed kernel fuction with $\int K = 1$, $h > 0$ is a smoothing factor, and $K_h(\cdot) = (1/h^d)K(\cdot/h)$. If h is chosen such that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, then it is well known (see Devroye and

Györfi [4]) that the estimate is universally consistent, that is, $\mathbb{E}||f_n - f|| \to 0$ for any density. Also, Devroye [3] shows that for any $\epsilon > 0$,

$$\mathbb{P}\{||f_n - f|| - \mathbb{E}||f_n - f|| \geq \epsilon\} \leq e^{-n\epsilon^2/2}.$$

Using these properties, it is easy to see that the testing method based on the kernel density estimate is consistent in the sense that the probability of error converges to zero exponentially for all $f \in \bigcup_{j=1}^{k} \overline{H}_j$. In order to show this suppose that $f \in \overline{H}_1$, and put

$$\epsilon = \min_{j>1} ||f - f^{(j)}|| - ||f - f^{(1)}||.$$

Then

$$\begin{aligned}
\mathbb{P}\{\text{error}\} &\leq \mathbb{P}\{\exists j > 1 : ||f_n - f^{(1)}|| \geq ||f_n - f^{(j)}||\} \\
&\leq (k-1) \max_{j>1} \mathbb{P}\{||f_n - f^{(1)}|| \geq ||f_n - f^{(j)}||\} \\
&\leq (k-1) \max_{j>1} \mathbb{P}\{||f_n - f|| + ||f - f^{(1)}|| \geq ||f - f^{(j)}|| - ||f_n - f||\} \\
&\leq (k-1)\mathbb{P}\{2||f_n - f|| \geq \epsilon\} \\
&= (k-1)\mathbb{P}\{||f_n - f|| - \mathbb{E}||f_n - f|| \geq \epsilon/2 - \mathbb{E}||f_n - f||\} \\
&\leq (k-1)e^{-n/2([\epsilon/2 - \mathbb{E}||f_n - f||]^+)^2},
\end{aligned}$$

where the last inequality follows from the above-mentioned inequality of Devroye [3]. The consistency of f_n assures that for a sufficiently large n, $\mathbb{E}||f_n - f|| < \epsilon/4$ and for such n, $\mathbb{P}\{\text{error}\} \leq (k-1)e^{-n\epsilon^2/32}$. However, since $\mathbb{E}||f_n - f||$ may tend to zero at an arbitrarily slow rate (see Devroye [2]), the error exponent is not uniform: it depends on f. It is known (see Barron [1], LeCam [12]) that for the hypotheses $\overline{H}_j$ it is impossible to construct a test with a uniform error exponent.

Acknowledgements. We thank the reviewers for drawing our attention to relevant literature.

References

[1] Barron A.R. (1989). Uniformly powerful goodness of fit tests. *Annals of Statistics*, 17:107–124.

[2] Devroye, L. (1983b). On arbitrary slow rates of global convergence in density estimation. *Zeitschrift für Wahscheinlichkeitstheorie und verwandte Gebiete*. 62:475-483.

[3] Devroye, L. (1991). Exponential inequalities in nonparametric estimation. in: *Nonparametric Functional Estimation*, ed. G. Roussas, pp. 31–44, NATO ASI Series, Dordrecht: Kluwer Academic Publishers.

[4] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the* L_1 *View*. Wiley, New York.

[5] Devroye, L. and Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimates. *Annals of Statistics*, 24:2499–2512.

[6] Devroye, L. and Lugosi, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity, and Yatracos classes. *Annals of Statistics*, 25:2626-2635.

[7] Devroye, L. and Lugosi, G. (2000). *Combinatorial Methods in Density Estimation*. Springer, New York, 2000.

[8] Györfi, L. and van der Meulen, E.C. (1991). A consistent goodness of fit test based on the total variation distance. in *Nonparametric Functional Estimation and Related Topics*, editor G. Roussas, pp. 631–646, Kluwer Academic Publishers, Dordrecht, 1991.

[9] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. Wiley, New York, 1986.

[10] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

[11] Huber, P.J. (1981) *Robust Statistics*. Wiley, New York, 1981.

[12] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1:38–53.

[13] Yatracos, Y.G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics* , 13:768–774.

[14] Zabin, S.M. and Wright, G.A. (1994). Nonparametric density estimation and detection in impulsive interference channels–Part II: Detectors. *IEEE Transactions on Communication*, 42:1698–1711.