ROBUST HYPOTHESIS TESTING FOR MODELING ERRORS

Gökhan Gül and Abdelhak M. Zoubir

Signal Processing Group, Technische Universitat Darmstadt
Merckstrasse 25, Darmstadt 64283, Germany.
{ggul,zoubir}@spg.tu-darmstadt.de

ABSTRACT

We propose a minimax robust hypothesis testing strategy between two composite hypotheses determined by the neighborhoods of two nominal distributions with respect to the squared Hellinger distance. The robust tests obtained are the nonlinearly transformed versions of the nominal likelihood ratios, whereas the least favorable densities are derived in three different regions. In two of them, they are scaled versions of the corresponding nominal densities and in the third region they form a composite version of the two nominal densities. The outcomes and implications of the proposed robust test are discussed through comparisons with the recent literature.

Index Terms— Detection, hypothesis testing, robustness

1. INTRODUCTION

Detection theory has been the key driver of many practical applications, such as radar, sonar, seismology, communications, or biomedicine. An event, for instance the existence or absence of a target is modeled statistically in terms of hypothesis testing. An optimum decision rule requires exact knowledge of the conditional densities under each hypothesis ($\mathcal{H}_0$ and $\mathcal{H}_1$). However for many practical applications, either the conditional densities are not completely known, possibly arising from physical considerations with a few unknown parameters [1], or can be affected by outliers such as impulsive noise [2], e.g., EEG signal contaminated by artifacts, i.e., high amplitude spikes.

In the case of exact modeling of the conditional densities, there exist some drawbacks. First, only some small deviations from the model assumptions, in other words a few bad observations (outliers), may upset the test statistics. On the other hand, the test procedure can never maintain a certain level of performance, therefore any possible performance degradation can not be foreseen. In such cases, the probability distributions are modeled to belong to the class of distributions, called the uncertainty class in the vicinity of some nominal distributions [3].

The main idea of robust hypothesis testing is to maximize the detection performance for the worst case densities determined from the uncertainty class. Accordingly, a certain level of detection is always maintained. This procedure is called minimax detection and regarded as conservative because it assumes no a-priori knowledge about the conditional densities.

One of the earliest works in this area has been presented by Huber in 1965 where he introduced the existence of least favorable densities as well as the robust version of the likelihood ratio test for the ϵ-contaminated class of distributions [2]. The robust test is achieved by clipping the nominal likelihood ratios and it provides a very elegant way of modeling outliers. However, as mentioned before, not all uncertainties are due to outliers; some of them result from a modeling mismatch.

In [4], the uncertainty classes have been constructed using the relative entropy constraint. The authors assign any density function, which is at least ϵ-close to the nominal density to the corresponding class. The motivation behind the use of the relative entropy as a distance is two-folds. First, it is a natural metric for model mismatch, and second, it forms a natural distance between statistical models [5]. This paper has two basic assumptions, one of which is the monotone increasing likelihood ratio constraint and the other one is the symmetry of the nominal densities. Monotonicity of the likelihood ratio is a basic requirement for mathematical tractability and can be circumvented in the testing stage. However, the symmetry constraint imposes a substantial restriction of the selection of the nominal densities. Moreover, the relative entropy is not a metric, neither symmetric nor satisfies the triangle inequality, and scales in $[0,\infty]$.

We propose an alternative approach for modeling errors, considering the squared Hellinger distance. It scales in $[0,1]$, which makes the choice of ϵ simpler, and it is a symmetric distance measure. Moreover, the design in this paper doesn't require the symmetry assumption between the nominal densities.

The organization of this paper is as follows. In the next section we present the problem setup by introducing the preliminaries, minimax decision rules and finally the derivation of the least favorable densities (LFDs), and the corresponding decision rules. In section 3, some examples and numerical results are provided whereas in the last section the paper is concluded.

2. PROBLEM FORMULATION

2.1. Preliminaries

Consider a binary hypothesis testing problem defined on a probability space $(\Omega, \mathscr{A}, P_i)$,

$$\begin{aligned}\mathcal{H}_0 &: Y \sim f_0(y)\\ \mathcal{H}_1 &: Y \sim f_1(y),\end{aligned} \qquad (1)$$

where $\Omega = \mathbb{R}$ and Y is a random variable which has a density $f_i(y)$ when $\mathcal{H}_i$ is true. The density function $f_i(y)$ models a phenomenon such as the existence or absence of a signal in white Gaussian noise. Given an observation y, a randomized decision rule $u(y) \in \Delta$ is a pointwise Bernoulli random variable with a success probability $\delta(y)$, where Δ stands for the set of all possible decision rules. A randomized decision rule can simply be specified by defining a function bounded in [0,1] on the real numbers. An optimum decision strategy, which minimizes the probability of error under both Bayesian

This work was supported by the LOEWE Priority Program Cocoon (http://www.cocoon.tu-darmstadt.de).

and Neyman-Pearson (NP) sense is the likelihood ratio test

$$L(y) = \frac{f_1(y)}{f_0(y)} \overset{u=1}{\underset{u=0}{\gtrless}} \gamma, \tag{2}$$

where γ is a constant threshold [6, pp. 65, 81]. The losses incurred from a chosen decision rule $u(y) \in \Delta$, given that $\mathcal{H}_0$ is true,

$$P_E^1(\delta, f_0) = \int_{\mathbb{R}} \delta(y) f_0(y) dy \tag{3}$$

and $\mathcal{H}_1$ is true,

$$P_E^2(\delta, f_1) = \int_{\mathbb{R}} (1 - \delta(y)) f_1(y) dy \tag{4}$$

are called false alarm and miss detection probabilities respectively. Eq. (2) also implies that no randomized decision rule can improve the Bayes or NP risk attained with a nonrandomized decision rule. Assuming (in Bayes' sense) that the costs of errors (false and miss detections) are the same and equal to one and the costs of type I and type II detections are zero, the minimum error detection can be written as

$$P_E(\delta, f_0, f_1) = P(\mathcal{H}_0) P_E^1(\delta, f_0) + P(\mathcal{H}_1) P_E^2(\delta, f_1). \tag{5}$$

In the next subsections, minimax rules, the existence of solutions to the minimax equations and the considered class of distributions are introduced.

2.2. Minimax Detection

Robust hypothesis testing differs from hypothesis testing which minimizes the error probability w.r.t. the nominal densities f_i via considering a set of densities $\mathcal{F}_i$ where each member of the set g_i is at least ϵ-close to f_i, $i = 0, 1$. We consider the squared Hellinger distance

$$S(g_i, f_i) = H^2(g_i, f_i) = \frac{1}{2} \int_{\mathbb{R}} \left(\sqrt{g_i(y)} - \sqrt{f_i(y)} \right)^2 \mathrm{d}y \tag{6}$$

to built the proximity sets,

$$\mathcal{F}_i = \{ g_i : S(g_i, f_i) \leq \epsilon_i \}. \tag{7}$$

Since the second derivative of $(\sqrt{g_i(y)} - \sqrt{f_i(y)})^2$ w.r.t. g is greater than zero, it is convex, which also implies that $\mathcal{F}_i$ is convex. Given the a-priori probabilities $P(\mathcal{H}_i)$, the probability of error, $P_E(\delta, f_0, f_1)$, is linear therefore both convex and concave in all three terms δ, f_0, f_1. Noting that $\mathcal{F}_0 \times \mathcal{F}_1$ as well as Δ are found to be both convex and compact [4], von Neumann's minimax theorem is applicable and we have

$$\begin{aligned} &\max_{(g_0, g_1) \in \mathcal{F}_0 \times \mathcal{F}_1} \min_{\delta \in \Delta} P_E(\delta, g_0, g_1) = \\ &\min_{\delta \in \Delta} \max_{(g_0, g_1) \in \mathcal{F}_0 \times \mathcal{F}_1} P_E(\delta, g_0, g_1), \end{aligned} \tag{8}$$

which also proves that there exists a unique solution to (8) with the so called least favorable densities $(\hat{g}_0, \hat{g}_1) \in \mathcal{F}_0 \times \mathcal{F}_1$ and the robust decision rule $\hat{\delta} \in \Delta$ [7]. The solution of (8) with $\{\hat{\delta}, (\hat{g}_0, \hat{g}_1)\}$ suggests a saddle point [4],

$$P_E(\delta, \hat{g}_0, \hat{g}_1) \geq P_E(\hat{\delta}, \hat{g}_0, \hat{g}_1) \geq P_E(\hat{\delta}, g_0, g_1). \tag{9}$$

2.3. Derivation of LFDs and Robust Decision Rules

A robust detection scheme is completely specified by the derivation of least favorable densities $(\hat{g}_0, \hat{g}_1)$ and a robust decision rule $\hat{\delta}$. From [2], we know that if a solution to (8) exists, it also solves

$$P_E^1(\hat{\delta}, g_0) \leq P_E^1(\hat{\delta}, \hat{g}_0) \leq P_E^1(\delta, \hat{g}_0), \tag{10}$$

and

$$P_E^2(\hat{\delta}, g_1) \leq P_E^2(\hat{\delta}, \hat{g}_1) \leq P_E^2(\delta, \hat{g}_1). \tag{11}$$

The least favorable densities should satisfy the first inequalities in (10) and (11). Therefore, we need to find some continuous function $g_i(y)$ such that

$$\hat{g}_0 = \arg\max_{g_0} P_E^1(\delta, g_0), \hat{g}_1 = \arg\max_{g_1} P_E^2(\delta, g_1) \tag{12}$$

subject to

$$g_i(y) > 0,\ \Upsilon(g_i) = \int_{\mathbb{R}} g_i(y) \mathrm{d}y = 1,\ g_i(y) \in \mathcal{F}_i,\ i = 0, 1 \tag{13}$$

This can be done by using Lagrange multipliers [8] because the optimization parameter g_i is convex in the constraints and concave in P_E^i, $i = 0, 1$. The positivity constraint will not be introduced explicitly as will be seen in the sequel the LFDs are always positive. Consider the following two Lagrangians $i = 0, 1$ which are coupled by $\hat{\delta}(y)$,

$$\mathcal{L}^i(g_i, \lambda_i, \mu_i) = P_E^i(\hat{\delta}, g_i) + \lambda_i(\epsilon_0 - S(g_i, f_i)) + \mu_i(1 - \Upsilon(g_i))). \tag{14}$$

The Lagrangian multipliers μ_i and λ_i are imposed such that g_i are densities and they belong to $g_i \in \mathcal{F}_i$. Eq. (14) can be explicitly rewritten for P_E^0 as

$$\begin{aligned} \mathcal{L}^0(g_0, \lambda_0, \mu_0) = &\int_{\mathbb{R}} \hat{\delta}(y) g_0(y) + \lambda_0 \epsilon_0 \\ &- \frac{\lambda_0}{2} \left(\sqrt{g_0(y)} - \sqrt{f_0(y)} \right)^2 + \mu_0 - \mu_0 g_0(y) \mathrm{d}y. \end{aligned} \tag{15}$$

Similarly to obtain $\mathcal{L}^1$, (15) should be manipulated with the following assignments: $\mu_0 := \mu_1$, $\lambda_0 := \lambda_1$, $\hat{\delta} := 1 - \hat{\delta}$. Taking the Gâteaux's derivative [8] of (15), we get

$$\int_{\mathbb{R}} [\hat{\delta}(y) + \frac{\lambda_0}{2\sqrt{g_0(y)}} \left(\sqrt{g_0(y)} - \sqrt{f_0(y)} \right) - \mu_0] z \mathrm{d}y \tag{16}$$

where z is an arbitrary function. Therefore the maximization in (8) reduces to the solution of

$$\hat{\delta}(y) + \frac{\lambda_0}{2\sqrt{g_0(y)}} \left(\sqrt{g_0(y)} - \sqrt{f_0(y)} \right) - \mu_0 = 0 \tag{17}$$

for which we get the LFD under $\mathcal{H}_0$ as

$$\hat{g}_0(y) = \frac{1}{\left(1 + 2 \left(\frac{\mu_0 - \hat{\delta}(y)}{\lambda_0} \right) \right)^2} f_0(y), \tag{18}$$

and similarly under $\mathcal{H}_1$ as

$$\hat{g}_1(y) = \frac{1}{\left(1 + 2 \left(\frac{\mu_1 - 1 + \hat{\delta}(y)}{\lambda_1} \right) \right)^2} f_1(y). \tag{19}$$

Clearly for any positive μ_i and λ_i, (18) and (19) are positive definite as claimed before. To minimize the objective function in (8), we have the likelihood ratio test between LFDs

$$\hat{L}(y) = \frac{\hat{g}_1(y)}{\hat{g}_0(y)} = \frac{\left(\frac{1}{2} + \left(\frac{\mu_0 - \hat{\delta}(y)}{\lambda_0}\right)\right)^2}{\left(\frac{1}{2} + \left(\frac{\mu_1 - 1 + \hat{\delta}(y)}{\lambda_1}\right)\right)^2} L(y), \tag{20}$$

and the decision rule is of the form

$$\hat{\delta}(y) = \begin{cases} 0, & \hat{L}(y) < \frac{P_0}{P_1} \\ \text{Arbitrary}, & \hat{L}(y) = \frac{P_0}{P_1} \\ 1, & \hat{L}(y) > \frac{P_0}{P_1} \end{cases}. \tag{21}$$

For the sake of simplicity, we assume that a-priori probabilities are equal and therefore the threshold in (21) can be taken as $P_0/P_1 = 1$. We use randomized decision rules instead of deterministic rules since it allows us to determine unique LFDs under $\mathcal{H}_0$ and $\mathcal{H}_1$ when $\hat{L}(y) = 1$. As it can be seen, (18) and (19) are functions of $\hat{\delta}(y)$, which is defined in (21). When $\hat{\delta}(y) = 0$ or $\hat{\delta}(y) = 1$, (the two cases in (21)), (18) and (19) have a simple scaled form of the nominal densities. When $\hat{L}(y) = 1$, we need to solve (20) in terms of $\hat{\delta}(y)$ and substitute the solution to either of (18) and (19). The solution of (20) leads us rewrite (21) as

$$\hat{\delta}(y) = \begin{cases} 0, & \hat{L}(y) < 1 \\ \frac{2\mu_0\lambda_1\sqrt{L(y)} + \lambda_0\left(2 - 2\mu_1 + \lambda_1(-1 + \sqrt{L(y)})\right)}{2\left(\lambda_0 + \lambda_1\sqrt{L(y)}\right)} & \hat{L}(y) = 1 \\ 1, & \hat{L}(y) > 1 \end{cases} \tag{22}$$

and define the least favorable distributions with respect to their densities;

$$\hat{g}_0(y) = \begin{cases} \frac{1}{4\left(\frac{1}{2} + \frac{\mu_0}{\lambda_0}\right)^2} f_0(y), & \hat{L}(y) < 1 \\ \frac{\left(\lambda_0\sqrt{f_0(y)} + \lambda_1\sqrt{f_1(y)}\right)^2}{(\lambda_0 + \lambda_1 + 2(\mu_0 + \mu_1 - 1))^2}, & \hat{L}(y) = 1 \\ \frac{1}{4\left(\frac{1}{2} + \frac{\mu_0 - 1}{\lambda_0}\right)^2} f_0(y), & \hat{L}(y) > 1 \end{cases} \tag{23}$$

and

$$\hat{g}_1(y) = \begin{cases} \frac{1}{4\left(\frac{1}{2} + \frac{\mu_1 - 1}{\lambda_1}\right)^2} f_1(y), & \hat{L}(y) < 1 \\ \frac{\left(\lambda_0\sqrt{f_0(y)} + \lambda_1\sqrt{f_1(y)}\right)^2}{(\lambda_0 + \lambda_1 + 2(\mu_0 + \mu_1 - 1))^2}, & \hat{L}(y) = 1 \\ \frac{1}{4\left(\frac{1}{2} + \frac{\mu_1}{\lambda_1}\right)^2} f_1(y), & \hat{L}(y) > 1 \end{cases} \cdot \tag{24}$$

Note that (20) has two roots and we consider the one given in (22). Because the other root leads to $g_i(y) = 0$ when $\lambda_0\sqrt{f_0(y)} = \lambda_1\sqrt{f_1(y)}$, and this contradicts with the constraints defined in (13). As mentioned earlier, we assume that the nominal likelihood ratio $L(y)$ is monotonically increasing. Therefore, there is a one-to-one correspondence between $y \in \mathbb{R}$ and $L(y)$, which in turns implies a one-to-one correspondence between $y \in \mathbb{R}$ and $\hat{L}(y)$ when $\hat{L}(y) \neq 1$ (c.f. (20)). To exploit this, we define two real numbers y_u and y_l, ($y_u > y_l$) corresponding to the upper $\hat{L}(y) > 1$ and lower limits $\hat{L}(y) < 1$ of LFDs. If we write the upper and lower inequalities explicitly using (20) and assign $\hat{\delta}(y) = 1$ and $\hat{\delta}(y) = 0$, respectively, we get the upper bound

$$y_u = L^{-1}\left[\left(\frac{\frac{1}{2} + \frac{\mu_1}{\lambda_1}}{\frac{1}{2} + \frac{\mu_0 - 1}{\lambda_0}}\right)^2\right] \tag{25}$$

and the lower bound

$$y_l = L^{-1}\left[\left(\frac{\frac{1}{2} + \frac{\mu_1 - 1}{\lambda_1}}{\frac{1}{2} + \frac{\mu_0}{\lambda_0}}\right)^2\right] \tag{26}$$

in terms of Lagrangian multipliers and the nominal likelihood ratio. Accordingly, the limits in equations (22)-(24) can be rearranged as $\hat{L}(y) > 1 \rightarrow y > y_u$, $\hat{L}(y) = 1 \rightarrow y_u > y > y_l$ and $\hat{L}(y) < 1 \rightarrow y < y_l$, if necessary. We note that the likelihood ratio $\hat{L}(y)$ maps to a measurable set on Ω while $L(y)$ maps to a single point when $\hat{L}(y) = 1$ and $L(y) = 1$, respectively. This is mainly where the robustness comes from. Due to modeling errors densities are known uncertainly around the point where $L(y) = 1$. To account for this, some amount of density under each hypothesis is delivered to the points at the neighborhood of $L(y) = 1$ as well as to the tails where the other hypothesis is preferred. Note that when $y_u > y > y_l$, both g_0 and g_1 are the same and no decision rule can provide a detection rate below 0.5. However, as we have shown earlier, there exists a unique decision rule when we allow randomized decision rules.

So far we obtained the LFDs and the robust decision rule in four parameters. In order to determine these parameters, we use two constraints imposed in the Lagrangian definition. Namely, we rewrite (6) and the second equation in (13) for (23) and (24). Accordingly, we obtain four nonlinear equations with four positive unknowns;

$$c_1\int_{-\infty}^{y_l} f_0(y)\mathrm{d}y + \int_{y_l}^{y_u} \Phi(y)\mathrm{d}y + c_2\int_{y_u}^{\infty} f_0(y)\mathrm{d}y = 1$$

$$c_3\int_{-\infty}^{y_l} f_1(y)\mathrm{d}y + \int_{y_l}^{y_u} \Phi(y)\mathrm{d}y + c_4\int_{y_u}^{\infty} f_1(y)\mathrm{d}y = 1$$

$$\begin{aligned} &\sqrt{c_1}\int_{-\infty}^{y_l} f_0(y)\mathrm{d}y + \int_{y_l}^{y_u} \sqrt{\Phi(y)f_0(y)}\mathrm{d}y + \sqrt{c_2}\int_{y_u}^{\infty} f_0(y)\mathrm{d}y \\ &= 1 - \epsilon_0 \\ &\sqrt{c_3}\int_{-\infty}^{y_l} f_1(y)\mathrm{d}y + \int_{y_l}^{y_u} \sqrt{\Phi(y)f_1(y)}\mathrm{d}y + \sqrt{c_4}\int_{y_u}^{\infty} f_1(y)\mathrm{d}y \\ &= 1 - \epsilon_1 \end{aligned} \tag{27}$$

where

$$c_1 = \frac{1}{4\left(\frac{1}{2} + \frac{\mu_0}{\lambda_0}\right)^2}, \; c_2 = \frac{1}{4\left(\frac{1}{2} + \frac{\mu_0 - 1}{\lambda_0}\right)^2}$$

$$c_3 = \frac{1}{4\left(\frac{1}{2} + \frac{\mu_1 - 1}{\lambda_1}\right)^2}, \; c_4 = \frac{1}{4\left(\frac{1}{2} + \frac{\mu_1}{\lambda_1}\right)^2}$$

and

$$\Phi(y) = \frac{\left(\lambda_0\sqrt{f_0(y)} + \lambda_1\sqrt{f_1(y)}\right)^2}{\left(\lambda_0 + \lambda_1 + 2\left(\mu_0 + \mu_1 - 1\right)\right)^2}.$$

3. EXAMPLES

In this section we perform some experiments to compare the results with the recent literature. In the first experiment, we consider the same example in [4]; $f_0 \sim \mathcal{N}(-1, 1)$ and $f_1 \sim \mathcal{N}(1, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ stands for the Gaussian density with mean μ and variance σ^2. Note that the symmetry condition $f_0(y) = f_1(-y)$ is satisfied. Figures 1 and 2 illustrate the LFDs for Hellinger and Kullback-Leibler distances when $\epsilon = \epsilon_0 = \epsilon_1$ and $y_u = 0.608$. We can see

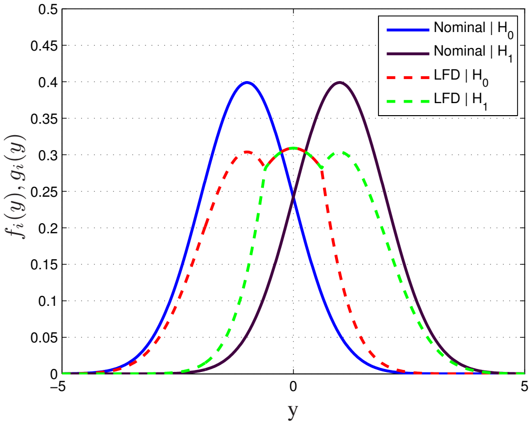

Fig. 1. LFDs based on Hellinger distance (symmetric)

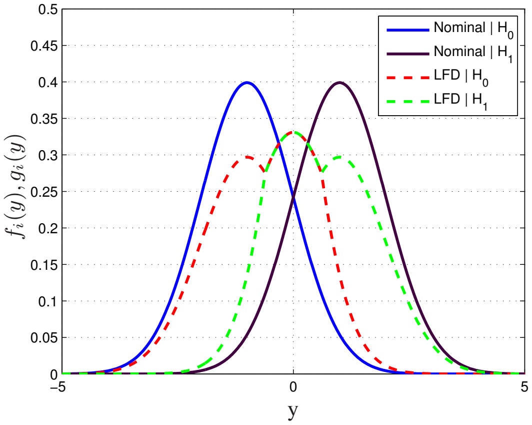

Fig. 2. LFDs based on relative entropy [4]

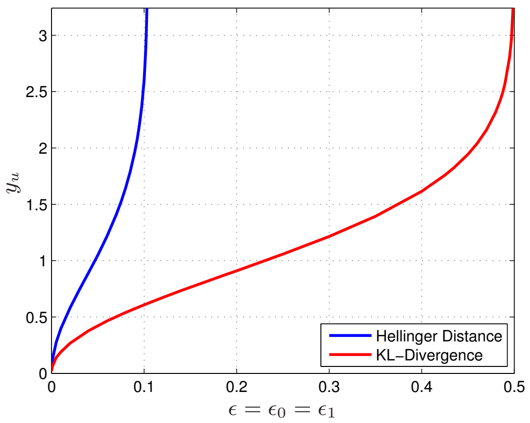

Fig. 3. Change in the overlapping region

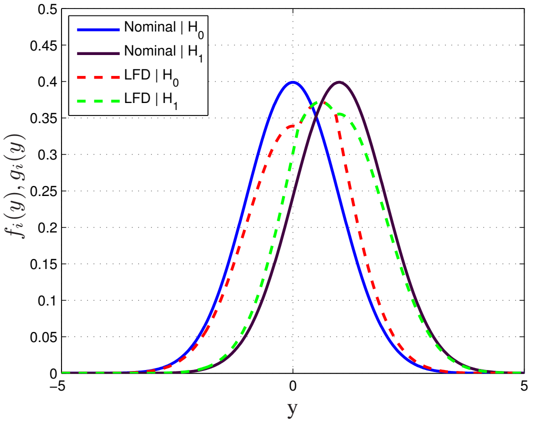

Fig. 4. LFDs based on Hellinger distance (non-symmetric)

that Hellinger distance delivers less amount of density to the overlapping regions compared to relative entropy. This suggests that the Hellinger distance is less conservative and thus provides higher detection rates for the worst case densities. In Fig. 3, we show that Hellinger distance is very sensitive to the small changes in ϵ compared to the relative entropy distance. In the last experiment, we consider $f_0 \sim \mathcal{N}(0,1)$, $f_1 \sim \mathcal{N}(1,1)$ for $\epsilon_0 = 0.005$ and $\epsilon_1 = 0.0025$. Note that the densities doesn't satisfy the symmetry constraint. In Figure 4, we can see that the clipping (overlapping) region has a non-symmetric shape w.r.t. $L(y) = 1$.

4. CONCLUSIONS

We have proposed a robust version of hypothesis testing, defining the conditional densities in the proximity of the nominal densities w.r.t. the squared Hellinger distance. The squared Hellinger distance, on the other hand, provides an alternative and elegant way of such a robust design since it scales in $[0,1]$ and requires no symmetry assumption to construct robust tests. The drawback of the Hellinger distance compared to the relative entropy is the computational complexity for densities satisfying the symmetry assumption. This complexity can be reduced taking into account the symmetry properties of the densities. The experimental results indicate that the Hellinger distance delivers less amount of density to the region where the nominal likelihood ratio is close to 1. This eventually implies reduced loss of performance due to robustness compared to the Kullback-Leibler distance. We have also justified that a little increase in ϵ results in a wider range of clipping region.

5. REFERENCES

[1] Bernard C. Levy, *Principles of Signal Detection and Parameter Estimation*, Springer Publishing Company, Incorporated, 1 edition, 2008.

[2] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, 1965.

[3] V. V. Veeravalli, T. Basar and H. V. Poor, "Minimax robust decentralized detection," *IEEE Trans. Inform. Theory*, vol. 40, pp. 35–40, Jan 1994.

[4] Bernard C. Levy, "Robust hypothesis testing with a relative entropy tolerance.," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, 2009.

[5] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical monographs*, Oxford University Press, 2000.

[6] S.M. Kay, *Fundamentals of statistical signal processing, Volume II: Detection theory*, vol. 7, Upper Saddle River (New Jersey), 1998.

[7] H. Tuy, *Minimax theorems revisited*, Acta Mathematica Vietnamica, 2004.

[8] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.