Minimax Robust Hypothesis Testing

Gökhan Gül, Student Member, IEEE, Abdelhak M. Zoubir, Fellow, IEEE,

Abstract—The minimax robust hypothesis testing problem for the case where the nominal probability distributions are subject to both modeling errors and outliers is studied in twofold. First, a robust hypothesis testing scheme based on a relative entropy distance is designed. This approach provides robustness with respect to modeling errors and is a generalization of a previous work proposed by Levy. Then, it is shown that this scheme can be combined with Huber’s robust test through a composite uncertainty class, for which the existence of a saddle value condition is also proven. The composite version of the robust hypothesis testing scheme as well as the individual robust tests are extended to fixed sample size and sequential probability ratio tests. The composite model is shown to extend to robust estimation problems as well. Simulation results are provided to validate the proposed assertions.

Index Terms—Detection, hypothesis testing, robustness, least favorable distributions, minimax optimization, sequential probability ratio test.

I. INTRODUCTION

The detection of the presence, or absence of an event with a specified accuracy is fundamental to statistical inference and binary hypothesis testing is the usual starting point. There are many applications, where binary hypothesis testing is used, for instance, radar, sonar, digital communications or seismology. A natural extension of binary hypothesis testing is multiple hypothesis testing, which builds a basis for classification and its importance is evident, for example, with pattern recognition. The necessity for statistical inference lies in the randomness that is inherent in the natural world such that received data, or signal, has an additive random component or, as in cognitive radio, must be modeled in a purely random manner. The degree of randomness in the received data usually turns out to be a metric of detection accuracy [1].

Formally, any real world example of binary decision making problem can be modeled by a binary hypothesis test, where under each hypothesis \( \mathcal{H}_j \), a received data \( y \in \mathbb{R} \) follows a particular probability distribution \( F_j, j \in \{0, 1\} \). Accordingly, the aim is to find a decision rule \( \delta \) which assigns each \( y \) either to \( \mathcal{H}_0 \) or \( \mathcal{H}_1 \), depending on a certain objective function, which can be, for instance, the error probability. An optimal decision rule \( \delta \) minimizes the objective function if \( y \) indeed follows \( F_j \) under \( \mathcal{H}_j, j \in \{0, 1\} \). However, this condition is too strict and often there are deviations from the model assumptions [2].

A traditional way of considering the deviations from the nominal distributions is via parametric modeling. Such parameters could be, for instance, the imprecisely known frequency of a receive signal or the unknown variance of a noise source. The shape of the probability distributions under each hypothesis is still assumed to be completely known. However, this assumption is invalid for various applications, for instance sonar, or cognitive radio. Obviously, in such cases, a parametric model is inappropriate, or if such a model is used, then severe performance degradation results.

The shortcomings of parametric modeling necessitate the use of non-parametric approaches. Such approaches are robust, are cheap to implement in practice, make (almost) no assumption on the nominal distributions and their performance is acceptable for a variety of detection problems [3]. However, compared to an optimum detector, their performance can be far away from being satisfactory, especially if there is some a priori knowledge available about the nominal distributions. Therefore, a more realistic approach should be tunable, depending on how much knowledge is available on the nominal distributions, how much robustness/performance trade-off is allowed as well as how complex the detector structure can be. In this context, robust minimax hypothesis testing falls between parametric and non-parametric detection; it coincides with parametric detection when the robustness parameters are chosen to be zero and it tends to a non-parametric test, a sign test, when the robustness parameters are chosen to be at maximum [4, p. 271].

A well known formulation of minimax hypothesis testing is based on building uncertainty sets \( \mathcal{F}_j \) for each hypothesis \( \mathcal{H}_j \), where \( \mathcal{F}_j \) are populated by all probability distributions \( G_j \), which are at least \( \epsilon_j \) close to the nominal distribution \( F_j \) with respect to some well defined distance \( D_j, j \in \{0, 1\} \). The choice of the parameters \( \epsilon_0 \) and \( \epsilon_1 \) determines the degree of robustness and they can vary with application. The eventual aim of the designer is to determine a pair of distributions \((G_0, G_1) \in \mathcal{F}_0 \times \mathcal{F}_1\) and a decision rule \( \delta \), such that a predefined performance measure is met, e.g. the bounded error probability. This type of optimization is called minimax optimization and the distributions solving this problem are called least favorable distributions (LFD)s.

In this research field, there are two main approaches: one of which was initiated by Huber [5] and the other by Dabak et al. [6] and Levy [7]. In Huber’s work, which was published as early as 1965, he proposed a robust version of the probability ratio test for \( \epsilon \)-contamination and total variation classes of distributions. He proved the existence of LFDs for both classes and showed that the resulting robust test was a censored/clipped version of the nominal likelihood ratio test. In a follow up work, he showed that the same conclusion could be made if the \( \epsilon \)-contamination model was extended to a larger class, which included five different distances as special cases [8]. A more general uncertainty class, called 2-alternating capacities, was proposed later by Huber and Strassen [9]. However, it was noted in [2] that the approach in [8] is more suitable for engineering applications due to its

G. Gül and A. M. Zoubir are with the Signal Processing Group, Institute of Telecommunications, Technische Universität Darmstadt, 64283, Darmstadt, Germany (e-mail: ggul@spg.tu-darmstadt.de; zoubir@spg.tu-darmstadt.de) Manuscript received April 19, 2005; revised January 11, 2007.
simplicity. In a recent work, an uncertainty class which allows the use of composite distances for robust hypothesis testing has been proposed [10]. The robust tests pioneered by Huber were designed for modeling outliers. More recent works by Dabak [6], and later by Levy [7], show that when the distance $D$ is chosen to be the relative entropy, the resulting robust tests are different from Huber’s robust test, depending on the choice of the objective function to be minimized. While Dabak’s approach minimizes the relative entropy between the LFDs and provides an asymptotically robust test, Levy’s robust test minimizes the type I and type II errors and provides a minimax robust test for a single sample. In [2], it was noted that the latter two robust tests are more appropriate for modeling errors instead of outliers. Recently, it has been shown that Levy’s robust test can be extended to distributed detection problems where the communication from sensors to the fusion center is constrained [11]. It has also been shown that considering the squared Hellinger distance instead of the relative entropy might provide a more flexible design [12], [13].

In this paper, a robust hypothesis testing scheme based on Kullback-Leibler (KL) divergence is proposed. The problem formulation doesn’t make any assumption about the choice of nominal distributions and, thus, it includes [7] as a special case. This robust scheme is then extended by use of a composite uncertainty set, which is built with respect to two different distances. The first distance models the misassumptions on the nominal distributions and the second distance models the outliers. It is proven that LFDs for this composite model exist and therefore a single test can be robust with respect to both modeling errors as well as outliers. Notice that this composite class is different from the one proposed in [10]. Finally, the designed robust tests are extended to fixed sample size and sequential probability ratio tests. It is also shown that the composite model can be extended to robust estimation problems.

The organization of this paper is as follows. In the following section, the LFDs and the robust decision rule are derived when the uncertainty sets are closed balls with respect to the KL divergence. The uniqueness and monotonicity properties of the LFDs are further proven. It is shown that the proposed model reduces to the model given in [7], when the nominal distributions are symmetric and the nominal likelihood ratio is monotone. For comparison reasons, the asymptotically robust test [6] is presented and the existence of LFDs is proven without considering the geometrical aspects of hypothesis testing. The implications of considering other distances to obtain the LFDs and the robust decision rules are also discussed. In Sec. [IV], the composite uncertainty set, which models both the outliers as well as the modeling errors, is introduced. When this model reduces to single robust tests, the density function of the log likelihood ratios is derived for performance evaluation as well as for asymptotic analysis. Similarly, the equations from which one can uniquely determine the maximum of the robustness parameters, above which a minimax robust test cannot be designed, are also derived. In Sec. [V], the robust methods are extended to fixed sample size tests. Especially, it is shown whether the robust tests maintain their LFD properties. The section is concluded with obtaining the limiting tests and the formulation of asymptotic analysis. In Sec. [VI], the sequential probability ratio test is robustified via replacing the nominal likelihood ratios by robust ones. It is investigated whether the LFD properties are preserved in general as well as asymptotically for both robustified sequential tests. In Sec. [VII], an extension of the composite uncertainty model for the design of robust estimation problems is briefly introduced. In Sec. [VIII], simulation results are presented and finally in Sec. [VIII], the paper is concluded.

II. ROBUST DETECTION FOR MODELING ERRORS

Let $(\Omega, \mathcal{F})$ be a measurable space with the probability measures $F_0$, $F_1$, $G_0$ and $G_1$ defined on it, which are absolutely continuous with respect to a dominating measure $\mu$, e.g. $\mu = F_0 + F_1 + G_0 + G_1$. Furthermore, let $f_0$, $f_1$, $g_0$ and $g_1$ be the density functions of the probability measures $F_0$, $F_1$, $G_0$ and $G_1$ with respect to $\mu$, respectively. Define the uncertainty classes

$$G_j = \{g_j : D(g_j, f_j) \leq \varepsilon_j\} \quad \varepsilon_j > 0$$

where every $g_j$ is at least $\varepsilon_j > 0$ close to the nominal density $f_j$, with respect to the KL-divergence i.e.

$$D(g_j, f_j) := \int \ln(g_j/f_j)g_j d\mu, \quad j \in \{0, 1\}.$$  

(2)

Now, consider the composite hypothesis testing problem

$$\mathcal{H}_0: Y \sim G_0$$

$$\mathcal{H}_1: Y \sim G_1$$

(3)

where $Y$ is a real-valued random variable (r.v.) on $\Omega$. Define a randomized decision rule (function) $\delta : \Delta \rightarrow [0, 1]$, where $\Delta$ stands for the set of all possible decision rules. Assume for the moment that $\varepsilon_0 = \varepsilon_1 = 0$. Then, the decision rule

$$\delta(y) = \begin{cases} 0, & l(y) < \rho \\ \kappa(y), & \rho \leq l(y) \leq \rho + 1 \\ 1, & l(y) > \rho + 1 \end{cases}$$

(4)

for some threshold $\rho = P(\mathcal{H}_0)/P(\mathcal{H}_1)$ and a function $\kappa : [0, 1] \rightarrow [0, 1]$, given the likelihood ratio $l(y) := f_1/f_0(y)$, is optimum in the sense that it minimizes the error probability both in the Bayes and the Neyman-Pearson sense and results in two types of errors: the false alarm probability

$$P^0_E(\delta, f_0) = \int_{\mathbb{R}} \delta f_0 d\mu$$

(5)

and the miss detection probability

$$P^1_E(\delta, f_1) = \int_{\mathbb{R}} (1 - \delta) f_1 d\mu.$$  

(6)

Accordingly, the minimum error probability is given by

$$P_E(\delta, f_0, f_1) = P(\mathcal{H}_0)P^0_E(\delta, f_0) + P(\mathcal{H}_1)P^1_E(\delta, f_1).$$  

(7)

Remark II.1. The sets $G_0$ and $G_1$ are not compact in the topology induced by the distance $D$. However, since $D$ is a convex function, $G_0$ and $G_1$ are convex sets. As a result $G_0 \times G_1$ is also convex. Given the a priori probabilities
$P(\mathcal{H}_0)$ and $P(\mathcal{H}_1)$, the probability of error $P_E(\delta, f_0, f_1)$ is continuous, real-valued and linear, and therefore both convex and concave in all three terms $\delta, f_0, f_1$. In general, the space of all randomized decision rules $\Delta = C^0(\mathbb{R}, [0, 1])$ is not compact. The compactness condition, however, is not required because the error minimizing decision rules are known to exist and to be the likelihood ratio test for all $(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1$. Let $\delta_1$ and $\delta_2$ be two decision functions chosen from $\Delta$. Then, simply for $\delta = \alpha \delta_1 + (1 - \alpha) \delta_2$, $0 \leq \alpha \leq 1$, we have $\delta \in \Delta$ and therefore $\Delta$ is convex. Note that any finitely supported quantization of $g_0$ and $g_1$ makes both $\mathcal{G}_0 \times \mathcal{G}_1$ and $\Delta$ compact with respect to the standard topology. This is a straightforward result of Heine-Borel theorem [13].

Remark [11] indicates that Sion’s minimax theorem [15] is applicable,

$$\sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} \min_{\delta \in \Delta} P_E(\delta, g_0, g_1) = \min_{\delta \in \Delta} \sup_{(g_0, g_1) \in \mathcal{G}_0 \times \mathcal{G}_1} P_E(\delta, g_0, g_1).$$

Hence, $P_E(\delta, g_0, g_1)$ possesses a saddle-value on $\Delta \times (\mathcal{G}_0 \times \mathcal{G}_1)$ with the least favorable densities $(\tilde{g}_0, \tilde{g}_1) \in \mathcal{G}_0 \times \mathcal{G}_1$ and the robust decision rule $\tilde{\delta} \in \Delta$, i.e., $\{\delta, (\tilde{g}_0, \tilde{g}_1)\}$, resulting from Eq. (8). Consequently

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

Since $P_E$ is distinct on $g_0$ and $g_1$, it follows that

$$P^0_E(\tilde{\delta}, g_0) \leq P^0_E(\tilde{\delta}, \hat{g}_0),$$

$$P^1_E(\tilde{\delta}, g_1) \leq P^1_E(\tilde{\delta}, \hat{g}_1).$$

**Theorem II.1.** Let $l_1$ and $l_u$ be two real numbers with $0 < l_1 \leq l_u < \infty$. Then, for

$$z(l_1, l_u) = \int_{l_1}^{l_u} f_1 d\mu + \int_{l_1 < l < l_u} (l_1^{-1} l)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1 d\mu + k(l_1, l_u) \int_{l > l_u} f_1 d\mu$$

and

$$k(l_1, l_u) = \frac{l_u - l_1}{l_u - l} f_0 d\mu$$

the least favorable densities

$$\hat{g}_0(y) = \begin{cases} \frac{l_1}{\ln(l_1)} f_0(y), & \text{if } l(y) < l_1 \\ \frac{l_1}{\ln(l_1)} \left(l_1^{-1} (1 - l(y) / l_1)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y), & \text{if } l_1 \leq l(y) \leq l_u \\ \frac{l_1}{\ln(l_1)} \left(l_1^{-1} l(y)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y), & \text{if } l(y) > l_u \end{cases}$$

$$\hat{g}_1(y) = \begin{cases} \frac{l_1}{\ln(l_1)} f_1(y), & \text{if } l(y) < l_1 \\ \frac{l_1}{\ln(l_1)} \left(l_1^{-1} l(y)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y), & \text{if } l_1 \leq l(y) \leq l_u \\ \frac{l_1}{\ln(l_1)} \left(l_1^{-1} (1 - l(y) / l_1)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y), & \text{if } l(y) > l_u \end{cases}$$

and the decision rule

$$\hat{\delta}(y) = \begin{cases} 0, & \text{if } l(y) < l_1 \\ \frac{l_1}{\ln(l_1)} \left(l_1^{-1} l(y)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y), & \text{if } l_1 \leq l(y) \leq l_u \\ 1, & \text{if } l(y) > l_u \end{cases}$$

which is equivalent to the robust likelihood ratio

$$\hat{l}(y) = \frac{l_1}{\ln(l_1)} \left(l_1^{-1} l(y)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1(y)$$

form the saddle value condition for Eq. (8). Furthermore the parameters $l_1$ and $l_u$ can be determined by solving

$$- \ln(z(l_1, l_u)) + \frac{1}{z(l_1, l_u)} \int_{l_1 < l < l_u} \frac{\ln(G(ll, lu))}{\ln(u/l_1)} f_1 d\mu$$

$$+ \int_{l_1 < l < l_u} (l_1^{-1} l)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} \ln \left(l_1^{-1} l \ln(l_1) \left(l_1^{-1} l(y)^{\frac{\ln(G(ll, lu))}{\ln(u/l_1)}} f_1 d\mu \right)$$

$$+ k(l_1, l_u) \int_{l_1 < l < l_u} f_1 d\mu = \epsilon_0$$

and

$$- \ln(z(l_1, l_u)) + \frac{1}{z(l_1, l_u)} \int_{l_1 < l < l_u} \frac{\ln(G(ll, lu))}{\ln(u/l_1)} f_1 d\mu$$

$$+ k(l_1, l_u) \int_{l_1 < l < l_u} f_1 d\mu = \epsilon_1.$$

**Proof:** The solution of the minimax non-linear optimization problem

$$\max_{g_0 \in \mathcal{G}_0 \times \mathcal{G}_1} P^0_E(\delta, g_j), \quad j \in \{0, 1\}$$

$$\text{s.t. } g_j > 0$$

$$\mathcal{Y}(g_j) = \int_{g_j} f_0 d\mu = 1, \quad j \in \{0, 1\}$$

$$\min_{\delta \in \Delta} P_E(\delta, \hat{g}_0, \hat{g}_1),$$

directly leads to the assertion. First, the maximization stage is solved by considering the Karush-Kuhn-Tucker (KKT) multipliers. The subsequent minimization and optimization stages complete the proof.

**A. Maximization stage**

Consider the Lagrangian

$$L^j(g_j, \lambda_j, \mu_j) = P^j_E(\hat{\delta}, g_j) + \lambda_j (\epsilon_j - \mathcal{Y}(g_j))$$

$$+ \mu_j (1 - \mathcal{Y}(g_j)), \quad j \in \{0, 1\}.$$ (19)

where $\mu_j, \lambda_j$ are the KKT multipliers which are imposed to satisfy the constraints. Since $L^j$ is concave in $g_j$, a globally optimum solution is guaranteed if the necessary KKT conditions are met [16]. Writing (19) explicitly for $P^0_E$, it follows that

$$L^0(\hat{g}_0, \lambda_0, \mu_0) = \int_{\mathbb{R}} \left[d\hat{g}_0 + \lambda_0 \epsilon_0 - \lambda_0 \ln \left(\frac{\hat{g}_0}{\int_{\mathbb{R}} \hat{g}_0 d\mu} + \mu_0 - \mu_0 \hat{g}_0\right)\right] d\mu$$

Imposing the first KKT condition (stationarity), through taking Gateaux’s derivative of Eq. (20) in the direction of $\psi$, yields

$$\int_{\mathbb{R}} [\delta - \lambda_0 \ln \left(\frac{\hat{g}_0}{\int_{\mathbb{R}} \hat{g}_0 d\mu} - \lambda_0 - \mu_0\right)] \psi d\mu,$$ (21)
which implies
\[ \delta - \lambda_0 \ln \frac{g_0}{f_0} - \lambda_0 - \mu_0 = 0, \quad (22) \]
since \( \psi \) is an arbitrary function. Hence, \( \hat{g}_0 \), and in a similar way \( \hat{g}_1 \) by solving (19) for \( P_E \), can be obtained. The results are
\[ \hat{g}_0 = c_1 \left( \frac{c_2}{c_1} \right)^{\delta} f_0, \quad \hat{g}_1 = c_3 \left( \frac{c_4}{c_3} \right)^{\delta} f_1 \quad (23) \]
where \( c_1 = \exp\left(-\frac{\lambda_0 + \mu_0}{\lambda_0} \right), \quad c_2 = \exp\left(-\frac{1+\lambda_0 + \mu_0}{\lambda_0} \right), \quad c_3 = \exp\left(-\frac{1}{\lambda_0 + \mu_0} \right) \) and \( c_4 = \exp\left(-\frac{1}{\lambda_0 + \mu_0} \right) \). This leads to the robust likelihood ratio
\[ \hat{l} = \left( \frac{c_3}{c_1} \right) e^{-\delta \ln \left( \frac{c_2}{c_1} \right)} l. \quad (24) \]

B. Minimization stage

The decision rule \( \delta \), which minimizes \( P_E \) for any \((g_0, g_1) \in G_0 \times G_1\), is known to be the likelihood ratio test [4]. Solving \( \hat{l} = 1 \) from Eq. (24) and rewriting Eq. (4) with \( \rho = 1 \) for \( \hat{l} \) yields
\[ \hat{\delta} = \begin{cases} 
0, & \hat{l} < 1 \\
\ln \left( \frac{\hat{l}}{\hat{l}} \right), & \hat{l} = 1 \\
1, & \hat{l} > 1
\end{cases} \quad (25) \]
Applying (25) to (23), the least favorable distributions with respect to their density functions are obtained as
\[ \hat{g}_0 = \begin{cases} 
c_{1} f_{0} l_{i}, \quad \hat{l} < 1 \\
c_{0} l_{i}, \quad \hat{l} = 1 \\
c_{2} f_{0}, \quad \hat{l} > 1
\end{cases} \quad (26) \]
where \( c_0 = \exp\left(\ln c_2 - \ln c_3 - \ln c_1 \ln c_4 \right) \). The unknown parameters can be obtained by imposing the constraints, or equivalently by solving the non-linear equations
\[ c_1 \int_{l_{i}} f_{0} d\mu + \int_{l_{i}} \Phi d\mu + c_2 \int_{l_{i}} f_{0} d\mu = 1 \]
\[ c_3 \int_{l_{i}} f_{1} d\mu + \int_{l_{i}} \Phi d\mu + c_4 \int_{l_{i}} f_{1} d\mu = 1 \]
\[ c_1 \int_{l_{i}} f_{1} d\mu + \int_{l_{i}} \Phi d\mu + c_2 \ln c_2 \int_{l_{i}} f_{0} d\mu = \varepsilon_0 \]
\[ c_3 \ln c_3 \int_{l_{i}} f_{1} d\mu + \int_{l_{i}} \Phi d\mu + c_4 \ln c_4 \int_{l_{i}} f_{1} d\mu = \varepsilon_1 \quad (27) \]
where \( \Phi = c_0 \exp\left(\ln c_1 - \ln c_2 - \ln c_3 + \ln c_4 \right) f_1 \). Note that the first two equations are required to make sure that \( \hat{g}_0 \) and \( \hat{g}_1 \) are density functions, i.e., they integrate to one and the other two equations are required to guarantee that \( \hat{g}_0 \in G_0 \) and \( \hat{g}_1 \in G_1 \).

C. Optimization stage

To complete the proof it is necessary to explain how \( \hat{\delta}, \hat{g}_0, \hat{g}_1 \) and the nonlinear equations can be represented in terms of \( l_i \) and \( l_u \). Let \( l_i = c_i/c_3 \) and \( l_u = c_2/c_4 \). Then, considering \( \hat{l} = \hat{g}_1/\hat{g}_0 \) from (26), it follows that
\[ R_1 := \{ y : l(y) < l_i \} \equiv \{ y : \hat{l}(y) < 1 \} \]
\[ R_2 := \{ y : l_i \leq l(y) \leq l_u \} \equiv \{ y : \hat{l}(y) = 1 \} \]
\[ R_3 := \{ y : l(y) > l_u \} \equiv \{ y : \hat{l}(y) > 1 \} \]
Rewriting the integrals with the new limits (over \( R_1, R_2, R_3 \)), using the substitutions \( c_1 := c_3 l_i \) and \( c_2 := c_4 l_u \), dividing both sides of the first two equations in (27) by \( c_3 \), and equating them to each other via \( 1/c_3 \) results in \( c_3 = k(l_i, l_u)c_3 \). Accordingly, it follows that
\[ \Phi = c_3 \left( \frac{l_{i}}{l_{u}} \right)^{\ln \left( k(l_i, l_u) \right)} f_1. \quad (28) \]
This allows the second equation in (27) to be written as \( c_3 := 1/z(l_i, l_u) \). Now, all constants \( c_1, c_2, c_3 \) and \( c_4 \) as well as \( \Phi \) are parameterized by \( l_i \) and \( l_u \). Thus, Eq. (24) can be rewritten as Eq. (15) and \( \hat{\delta}, \hat{g}_0, \hat{g}_1 \) as given in Theorem II.1. Finally, the last two equations of (27) reduce to (16) and (17). This completes the proof.

D. Monotonicity of the relative entropy

In the sequel it is shown that ordering in likelihood ratios implies ordering in KL-divergence. This explains the monotonic behavior of LFDs for increasing robustness parameters given that \( l \) is monotone. The theory that will be presented will also be used in the next sections.

**Proposition II.2.** Let \( F \) and \( G \) be two probability measures on \((\Omega, \mathcal{A})\) with \( \partial F / \partial G \) a non-decreasing function. Then, \( G(y) \geq F(y) \) for all \( y \in \mathbb{R} \).

**Proof:** Due to a special case of the Fortuin-Kasteleyn-Ginibre (FKG) inequality, for any random variable \( X \) and any two positive increasing functions \( \phi, \psi \) we have \( \mathrm{E} \phi(X) \psi(X) \geq \mathrm{E} \phi(X) \mathrm{E} \psi(X) \). Applying this to \( X \) distributed according to \( G \) and the functions \( \phi := 1_{[r, +\infty)} \), where \( 1_{[r, \infty)} \) is the indicator function, and \( \psi := dF/dG \), we get \( G(y) \geq F(y) \) for all \( y \in \mathbb{R} \).

**Remark II.2.** Let \( X \) and \( Y \) be two random variables defined on the same probability space \((\Omega, \mathcal{A})\), having continuous cumulative distribution functions \( F \) and \( G \), respectively. \( X \) is called stochastically larger than \( Y \), i.e., \( X \succ_{ST} Y \), if \( G(y) \geq F(y) \) for all \( y \).

**Corollary II.3.** For every non-decreasing function \( \phi, X \succ_{ST} Y \iff \phi(X) \succ_{ST} \phi(Y), \) hence \( X \succ_{ST} Y \iff \mathbb{E}[\phi(X)] \geq \mathbb{E}[\phi(Y)] \)

**Proof of Corollary II.3** is simple and can be found for example in [17].

**Theorem II.4.** Let \( X_0, Y_0, X_1, \) and \( Y_1 \) be four continuous random variables defined on \((\Omega, \mathcal{A})\) and having distinct densities \( f_0, g_0, f_1, \) and \( g_1 \), respectively, with \( f_1/g_1, g_1/g_0, \) and \( g_0/f_0 \), all being non-decreasing functions. Then,
\[ D(f_1, f_0) > D(g_1, g_0) \quad \text{and} \quad D(f_0, f_1) > D(g_0, g_1) \quad (29) \]
Proof: By Prop. [11.2] and Remark [11.1], we have $Y_1 \sim_{ST} Y_0$ and $Y_0 \sim_{ST} X_0$ since $g_1/g_0$, and $g_0/f_0$ are non-decreasing functions. Increasing $f_1/g_1$ and $g_1/g_0$ implies increasing $f_1/f_0$ and using Corollary [11.3] and denoting $\phi(Y) = \ln g_0/f_1(Y)$, we have $E_{X_0}[\phi(Y)] \geq E_{Y_0}[\phi(Y)]$. Hence, the identity $D(f_0, f_1) = E_{X_0}[\phi(Y)] + D(f_0, g_0)$, together with $E_{X_0}[\phi(Y)] \geq E_{Y_0}[\phi(Y)]$, results in $D(f_0, f_1) \geq D(f_0, g_0) + D(g_0, f_1) \implies D(f_0, f_1) > D(g_0, f_1)$. It is well known that $E_{Y_1} \left[ \ln \frac{g_1(Y)}{f_1(Y)} \right] > 0$. (30)

Again, using the Corollary [11.3] and denoting $\psi(Y) = \ln f_1/g_1(Y)$, we have $E_{Y_1}[\psi(Y)] \geq E_{X_0}[\psi(Y)]$, which implies $-E_{Y_1}[\psi(Y)] > 0$ in comparison with $-E_{Y_1}[\psi(Y)] > 0$. We conclude that $D(f_0, f_1) > D(g_0, f_1)$ together with $-E_{Y_1}[\psi(Y)] > 0$ implies $D(f_0, f_1) > D(g_0, f_1)$. The proof for the case $D(f_1, f_0) > D(g_1, g_0)$ is similar and is omitted.

E. Symmetric density functions

Depending on the extra constraints imposed on the nominal probability distributions, the equations that need to be solved to determine the parameters of the LFDs can be simplified. Assume $f_0(y) = f_1(-y)$ for all $y \in \mathbb{R}$ and $\varepsilon = \varepsilon_0 = \varepsilon_1$. This implies $l_u = 1/l_u$. With this assumption Eq. (16) and Eq. (17) reduce to

$$-\ln(z(l_u)) - \frac{1}{z(l_u)} \left( l_u^{-1/2} \ln l_u \int_{1<l<u} \sqrt{f_0 f_1} d\mu \right) + l_u^{-1} \ln l_u \int_{l>u} f_1 d\mu = \varepsilon$$

(31)

where

$$z(l_u) = \int_{1<l<u} f_1 d\mu + 2l_u^{-1/2} \int_{1<l<u} \sqrt{f_0 f_1} d\mu + l_u^{-1} \int_{l>u} f_1 d\mu.$$  

(32)

The symmetry condition also implies $l(y) = 1/l(-y)$ and $l(y) = 1/l(-y)$ for all $y$. Accordingly, it follows that $k(l, l_u) = l_u^{-1}$ and $\bar{g}_0(y) = g_1(-y)/y$. Notice that if $l$ is monotone, Eq. (31) can be redefined in terms of $y_a$ by $l_u = l(y_a), \{l > l_u\} \equiv (y_a, \infty)$ and due to symmetry $1/l_u = l(-y_a), \{l < l_u\} = (-\infty, -y_a)$. This proves that Theorem [11.1] is a generalization of the results of [7].

F. Asymptotically robust hypothesis test

So far, the problem of minimax robust hypothesis testing, for the case where the objective function to maximize was the error probability, has been studied. For the same uncertainty model (1), Dabak and Johnson proposed a geometrically based robust detection scheme much earlier than [7]. From [2, p.254], it is also known that the work of Dabak can be recreated by considering the same minimax optimization problem that has been introduced, see [15], but changing the objective functions $P_E$ and $P_E$ to $-D(g_0, g_1)$ and $-D(g_1, g_0)$. Here, $D$ is again the relative entropy and $(\bar{g}_0, \bar{g}_1)$ are the least favorable densities,

$$\bar{g}_0(y) = \frac{w(y; u)}{k(u)}, \quad \bar{g}_1(y) = \frac{w(y; 1 - v)}{k(1 - v)}$$  

(33)

where $u, v$ are parameters to be determined such that

$$D(\bar{g}_0, f_0) = \varepsilon_0, \quad D(\bar{g}_1, f_1) = \varepsilon_1.$$  

(34)

Again by [2], the fixed sample size test in the log domain

$$\frac{1}{n} \sum_{i=1}^{n} \ln (l(y_i)) \leq \frac{\ln \left( \frac{k(1-v)}{k(u)} \right)}{1 - (u + v)}$$

(35)

is still a likelihood ratio test, but with a modified threshold ($\gamma = 1$). The following proposition and the proof show that $\bar{g}_0$ and $\bar{g}_1$ are indeed LFDs without consideration of the geometrical aspects of hypothesis testing.

Proposition II.5. The pair of density functions $\bar{g}_0$ and $\bar{g}_1$ satisfy

$$\bar{g}_0 = \max_{g_0 \in G_0} E_{g_0} \ln \left( \frac{\bar{g}_1}{\bar{g}_0} \right), \quad \bar{g}_1 = \min_{g_1 \in G_1} E_{g_1} \ln \left( \frac{\bar{g}_1}{\bar{g}_0} \right).$$  

(36)

**Proof:** Consider the Lagrangian function defined in [19], where the objective functions $P_E$ and $P_E$ are replaced by $E_{g_0} \ln (\bar{g}_1/g_0)$ and $E_{g_1} \ln (\bar{g}_1/g_0)$. Then, following similar steps to [20]–[23], it can be shown that $g_0$ and $g_1$ have the same parametric forms as given in (35). The equations in (34) are convex [2], hence their solution is unique. Since $(\bar{g}_0, \bar{g}_1)$ must satisfy (34) with the same $(\varepsilon_0, \varepsilon_1)$ that $(\bar{g}_0, \bar{g}_1)$ must satisfy, we have $\bar{g}_0 = \bar{g}_0$ and $\bar{g}_1 = \bar{g}_1$. 

Note that $\bar{g}_0$ and $\bar{g}_1$ are denoted as least favorable densities only in the sense that they are solutions to the equations in (36). In the sequel, the statistical test based on the likelihood ratio $\bar{g}_1/\bar{g}_0$ will be denoted as the (a)-test. The property defined by (36) will be used in the next sections.

G. Other distances

The distance $D$ can be chosen in various ways based on mathematical tractability or the practical application [18]. Symmetric distances are preferable due to their nice properties; for instance, the symmetric version of the relative entropy $D(f_0, f_1) + D(f_1, f_0)$. However, this distance does not yield an analytic expression for the LFDs and the decision rule as

$$l(y) = W(e^{z_1 \delta(y) + z_1}) - W(e^{z_2 \delta(y) + z_2}) + z_3 \delta(y)$$

needs to be solved to obtain the decision rule $\delta(y)$ for $\bar{f} = 1$, where $z_1$, $z_2$, and $z_3$ are constants and $W$ is the Lambert $W$-function. Symmetrized $\chi^2$ distance, i.e. $\chi^2(f_0, f_1) + \chi^2(f_1, f_0)$, is another example where the LFDs can be obtained analytically. However, the relation between $y_a$ and $l_u$, and similarly between $l_1$ and $y_1$, cannot be obtained analytically. Another example for a symmetric distance is the
squared Hellinger distance. This distance is more appealing as it scales in \([0, 1]\) and it is mathematically tractable \([12, 13]\).

For various robust tests, including the relative entropy distance, the \(\chi^2\) distance and the squared Hellinger distance, the likelihood ratio test is given by \([15]\). For the symmetrized \(\chi^2\) distance, however, the test is slightly different as \(l/l\) is not a constant function for \(\delta = 0\) and \(\delta(y) = 1\), c.f. Sec. \(\[\text{VII}\]\). In general, designing a robust test is equivalent to determining \(\hat{l} = \psi(f_0, f_1)\) for some suitable functional \(\psi\) which accounts for the unmodeled uncertainties by the nominal model while maintaining the detection performance above a certain threshold.

### III. Robust Detection for the Composite Uncertainty Model

Minimax robust tests, which are designed based on a neighborhood set, where every probability measure belonging to the set is absolutely continuous with respect to the nominal distribution, e.g. \([1, 13]\), are more suitable for modeling errors than the tests designed based on a neighborhood set, where not all distributions are absolutely continuous with respect to the nominals e.g. \([5]\); see \([6]\) and \([7]\). In many practical applications, however, both types of uncertainties, namely both modeling errors as well as outliers can occur and a reasonable approach is to build a single test which is uniformly minimax robust. This can be done by combining one of Huber’s clipped likelihood ratio tests \([9]\) with a robust test which is more suitable for modeling errors. The following proposition explains how this can be done.

**Proposition III.1.** Let the inner uncertainty set be the extended version of \([4]\), i.e.

\[
G_j = \{g_j : D(g_j, f_j) \leq \varepsilon_j\}, \quad j \in \{0, 1\}
\]

where \(D\) is a convex distance (possibly different for each hypothesis), \(0 \leq \varepsilon_j < 1\) are some numbers and \(l = f_1/f_0\) is a monotone increasing function. Assume that there exist \(\hat{g}_0 \in G_0\) and \(\hat{g}_1\) in \(G_1\) corresponding to probability measures \(G_0\) and \(G_1\), respectively, such that

\[
G_0[\hat{g}_1/\hat{g}_0 \leq t] \geq \hat{G}_0[\hat{g}_1/\hat{g}_0 \leq t] \quad \forall t \in \mathbb{R}, \forall g_0 \in G_0
\]

\[
G_1[\hat{g}_1/\hat{g}_0 \leq t] \leq \hat{G}_1[\hat{g}_1/\hat{g}_0 \leq t] \quad \forall t \in \mathbb{R}, \forall g_1 \in G_1.
\]

**Define the composite uncertainty sets**

\[
\mathcal{F}_j = \{Q_j | Q_j = (1-\varepsilon_j)G_j + \varepsilon_j H_j, H_j \in \Xi, g_j \in G_j\}, \quad j \in \{0, 1\}
\]

where \(\Xi\) is the set of all probability measures on \((\Omega, \mathcal{A})\) and \(0 < \varepsilon_0, \varepsilon_1 < 1\). Then, there exist a pair of LFDs, \((\hat{Q}_0, \hat{Q}_1)\) which satisfy the saddle value condition

\[
\hat{Q}_0[\hat{g}_1/\hat{g}_0 \leq t] \geq \hat{Q}_0[\hat{g}_1/\hat{g}_0 \leq t] \quad \forall t \in \mathbb{R}, \forall Q_0 \in \mathcal{F}_0
\]

\[
\hat{Q}_1[\hat{g}_1/\hat{g}_0 \leq t] \leq \hat{Q}_1[\hat{g}_1/\hat{g}_0 \leq t] \quad \forall t \in \mathbb{R}, \forall Q_0 \in \mathcal{F}_0.
\]

if \(\varepsilon_j\) and \(\varepsilon_j\) are small enough, i.e., \(\mathcal{F}_0\) and \(\mathcal{F}_1\) do not overlap, where \(\hat{q}_0\) and \(\hat{q}_1\) are the least favorable densities

\[
\hat{q}_0(y) = (1-\varepsilon_0)\hat{g}_0(y) \quad \text{for} \quad \hat{g}_1(y)/\hat{g}_0(y) < c_u
\]

\[
= (1/c_u)(1-\varepsilon_0)\hat{g}_1(y) \quad \text{for} \quad \hat{g}_1(y)/\hat{g}_0(y) \geq c_u
\]

\[
\hat{q}_1(y) = (1-\varepsilon_1)\hat{g}_1(y) \quad \text{for} \quad \hat{g}_1(y)/\hat{g}_0(y) > c_t
\]

\[
= \varepsilon_1(1-\varepsilon_1)\hat{g}_0(y) \quad \text{for} \quad \hat{g}_1(y)/\hat{g}_0(y) \leq c_t
\]

(41)

**corresponding to \(\hat{Q}_0\) and \(\hat{Q}_1\), respectively.**

**Proof:** The proof follows directly from the definition of the uncertainty sets

\[
\mathcal{F}_0 = \{Q_0 : Q_0[Y < y] \geq (1-\varepsilon_0)G_0[Y < y] - \nu_0\}
\]

\[
\mathcal{F}_1 = \{Q_1 : 1 - Q_1[Y < y] \geq (1-\varepsilon_1)(1 - G_1[Y < y]) - \nu_1\}
\]

(42)

with \(\nu_0 = \nu_1 = 0\) for \(\varepsilon\)-contamination neighborhood \([8]\) and the stochastic ordering defined by Corollary \([13]\). Only the first inequality in \((40)\) can be proven using the same line of arguments. Let \(b = (1-\varepsilon_1)/(1-\varepsilon_0)\). Then, for every \(t > bc_u\) and \(Q_0 \in \mathcal{F}_0\), the event \(\{\hat{g}_1/\hat{g}_0 \leq t\}\) has full probability and for every \(t < bc_t\) and \(Q_0 \in \mathcal{F}_0\), the event \(E\) has null probability. Hence, \((40)\) is trivially true for these cases. For \(bc_t < t < bc_u\), assume that the likelihood ratio \(\hat{g}_1/\hat{g}_0\) is non-decreasing, which is true when \(t\) is monotone and the distance is either one of Huber’s distances \([4, p. 271]\) or any distance with the likelihood ratio given by Eq. \([15]\) or in general a distance which results in a non-decreasing \(l = \hat{g}_1/\hat{g}_0\) for monotone \(l\). Then, by Corollary \([13]\) it follows that \(\hat{G}_0[Y \leq t] \geq \hat{G}_0[Y \leq t]\) for all \(t \in \hat{l}^{-1}(y)\). Let \(\hat{Q}_0[Y \leq \hat{t}] := (1-\varepsilon_0)\hat{G}_0[Y \leq \hat{t}]\). Obviously \(\hat{Q}_0[Y \leq \hat{t}] \geq \hat{Q}_0[Y \leq \hat{t}]\) for all \(bc_t < t < bc_u\) and \(Q_0 \in \mathcal{F}_0\). Note that for non-decreasing \(\hat{g}_1/\hat{g}_0\), \(\hat{g}_1/\hat{g}_0\) is also non-decreasing. Hence, again by Corollary \([13]\), we get \(\hat{Q}_0[\hat{g}_1/\hat{g}_0 \leq \hat{t}] \geq \hat{Q}_0[\hat{g}_1/\hat{g}_0 \leq \hat{t}]\) for all \(t\) and \(Q_0 \in \mathcal{F}_0\) as claimed.

The proof is independent of the choice of \(D\) as long as the LFDs exist. When \(D\) is the relative entropy, it follows that

\[
\hat{G}_0[\hat{g}_1/\hat{g}_0 \leq \rho] = \int_{[\hat{g}_1/\hat{g}_0 < \rho]} \hat{g}_0d\mu + \int_{[\hat{g}_1/\hat{g}_0 = \rho]} \hat{\delta}g_0d\mu
\]

\[
\leq \int_{[\hat{g}_1/\hat{g}_0 < \rho]} \hat{g}_0d\mu + \int_{[\hat{g}_1/\hat{g}_0 = \rho]} \hat{\delta}g_0d\mu
\]

\[
= G_0[\hat{g}_1/\hat{g}_0 \leq \rho]
\]

(43)

and in a similar way \(\hat{G}_1[\hat{g}_1/\hat{g}_0 \leq \rho] \leq \hat{G}_1[\hat{g}_1/\hat{g}_0 \leq \rho]\). This proves that the uncertainty sets based on \(\varepsilon\)-contamination model and the relative entropy can be combined into a composite uncertainty set \([3]\) which accepts LFDs, \(Q_0\) and \(Q_1\) satisfying \((40)\). Clearly, the same conclusions hold when \(\nu_0\) and \(\nu_1\) are non-zero. This includes the total variation distance as a special case with \(\varepsilon_0 = \varepsilon_1 = 0\). Note that Prop. \([\text{VII}]\) is general for all thresholds. However, when the inner uncertainty set is the KL-divergence, the decision rule \(\delta\) must be used to guarantee minimax robustness. For a comparison, one can see that the composite model proposed in \([10]\) is robust only against outliers, with some flexibility, while the composite model proposed in this work is robust against both modeling errors as well as outliers. The LFDs, corresponding to the
composite model based on the relative entropy distance, can also be obtained as
\[ \hat{q}_0(y) = \begin{cases} \frac{(1-\epsilon_0)k}{z(t,\epsilon_0)} f_0(y), & l(y) > l_0 \\ \frac{(1-\epsilon_0)k}{z(t,\epsilon_0)} f_1(y), & l_0 \leq l(y) \leq l_u \\ \frac{(1-\epsilon_0)(1-k)}{z(t,\epsilon_0)} f_1(y), & c_l \leq l(y) < c_u \\ \frac{b_c(1-\epsilon_0)}{z(t,\epsilon_0)} f_0(y), & c_l \leq l(y) < c_u \\ \frac{b_c}{z(t,\epsilon_0)} f_0(y), & c_u \leq l(y) < c_l \\ \frac{b_c}{z(t,\epsilon_0)} f_1(y), & c_u \leq l(y) < c_u \\ \frac{b_u}{z(t,\epsilon_0)} f_0(y), & l(y) \geq c_u \\ \frac{b_u}{z(t,\epsilon_0)} f_1(y), & l(y) \geq c_u \end{cases} \]

with the corresponding likelihood ratio
\[ \hat{l}(y) = \begin{cases} bc_1, & l(y) \leq c_l \\ \frac{b_c}{z(t,\epsilon_0)l}, & c_l \leq l(y) < l_i \\ b_i, & l_i \leq l(y) \leq l_u \\ \frac{b_i}{z(t,\epsilon_0)}, & l_u \leq l(y) < c_u \\ \frac{b_c}{z(t,\epsilon_0)} l, & c_u \leq l(y) \leq c_u \end{cases} \]

The choice of D can be adjusted depending on the application. For instance symmetrized $\chi^2$ distance can be preferred if the tail structure is expected to be roughly preserved. It is also not difficult to see that for variety of distances, (44) remains the same. However, special care should be taken for the choice of b, since it is equivalent to $\rho$. In the sequel, D will be assumed to be KL-divergence with the LFDs given by (13) unless mentioned otherwise. For example the parameters $\epsilon_0 = \epsilon_1 = 0$ indicate a pure KL-divergence uncertainty set with the corresponding LFDs denoted by $\hat{Q}_0 := \hat{G}_0$ and $\hat{Q}_1 := \hat{G}_1$. In the following, the corresponding test $d\hat{G}_1/d\hat{G}_0$ will be denoted as the (m)-test and similarly, the minimax robust test for $\epsilon_0 = \epsilon_1 = 0$ will be denoted as the (h)-test and the composite test will be denoted as the (c)-test.

A. Distribution of the log-likelihood ratios of LFDs

In order to gain further insights about the minimax robust tests and to evaluate their performance, it is desirable to have the density function of the log likelihood ratio of the LFDs, i.e., $h^*_i \sim \ln \hat{q}_i/\hat{q}_0(Y)$, when $Y \sim \hat{Q}_j$, as a function of the density function of the log likelihood ratio of the nominal distributions $h_j \sim \ln f_j/f_0(Y)$ when $Y \sim F_j$, $j \in \{0,1\}$. Then, for the (h)-test, it follows that
\[ h^*_i(x) = r^*_i \delta_x(x - \ln (bc_i)) + (1 - \epsilon_i) h_i(x - \ln b) \mathbb{1}_{\{\ln (bc_i) < x < \ln (bc_u)\}} + r^*_i \delta_x(x - \ln (bc_u)), \quad i \in \{0,1\} \]

where $\delta_x$ is a dirac delta function and
\[ r^*_0 = (1 - \epsilon_0) F_0[f_1/f_0 \leq c_l], \quad r^*_1 = \frac{1}{c_u}(1 - \epsilon_0) F_1[f_1/f_0 \geq c_u] \]
\[ r^*_0 = c_l(1 - \epsilon_1) F_0[f_1/f_0 \leq c_l], \quad r^*_1 = (1 - \epsilon_1) F_1[f_1/f_0 \geq c_u] \]

Similarly, for the (m)-test,
\[ h^*_0(x) = \frac{l_i}{z(l_i, l_u)} h_0(x + \ln l_i) \mathbb{1}_{x < 0} + \frac{r}{z(l_i, l_u)} \delta_x(x) + \frac{lu k(l_i, l_u)}{z(l_i, l_u)} h_0(x + \ln l_u) \mathbb{1}_{x > 0} \]
\[ h^*_1(x) = \frac{1}{z(l_i, l_u)} h_1(x + \ln l_i) \mathbb{1}_{x < 0} + \frac{r}{z(l_i, l_u)} \delta_x(x) + \frac{k(l_i, l_u)}{z(l_i, l_u)} h_1(x + \ln l_u) \mathbb{1}_{x > 0} \]

where
\[ r = \int_{l_i < l < l_u} \left( l^{-1} \right)^{\mathbb{1}_{\{\ln(l_i)/l_u\}}} f_1 d\mu. \]

It can be seen that Huber’s test ((h)-test) creates two point masses at the clipping thresholds $(\ln(b_c), \ln(b_u))$ and between them the density of the log-likelihood ratio of the nominal distributions is shifted by $\ln b$. The robust test based on modeling errors ((m)-test), on the other hand, shifts the density of the log-likelihood ratio of the nominal distributions $(h_0, h_1)$ by $|\ln l_i|$ to the right and adds another part of the same density, which is shifted by $\ln l_u$, to the left. The total loss of area due to the shifting is stacked as a point mass at $x = 0$.

The equations (46) and (48) are of particular importance, first in calculating the false alarm and miss detection probabilities $\int_{-\infty}^{\infty} h_0^*(x) dx$ and $\int_{-\infty}^{\infty} h_1^*(x) dx$, respectively, and second in finding the approximate distribution of the test statistic $S_n = \sum_{i=1}^n \ln l(Y_i)$ for $n$ independent r.v.s $Y_1, Y_2, \ldots, Y_n$, in terms of nominal distributions. However, to calculate the false alarm and miss detection probabilities, the factor of randomization, $\delta$ in Eq. (48), needs to be taken into account. That is, the contribution of the point mass at $x = 0$ to the false alarm and miss detection probabilities needs to be determined.

B. Limiting robustness parameters for the (m)-test

The composite hypotheses start overlapping when the LFDs become identical. For the (m)-test, this occurs when $\mathcal{R}_1$ and $\mathcal{R}_3$ are empty sets. Let $u = 1 + \ln(k(l_i, l_u))/\ln(l_u/l_i)$, $w(y; u) = f_1(y)^u f_0(y)^{1-u}$ and $k(u) = \int_{\mathbb{R}} w(y; u) dy$. Then, equations (16) and (17) reduce to
\[ \varepsilon_j(u) = -\ln k(u) + \frac{u - j}{k(u)} \int_{\mathbb{R}} w(y; u) \ln l(y) dy, \quad j \in \{0,1\} \]

Proposition III.2. $\varepsilon_0$ is monotone increasing in $u$ and $\varepsilon_1$ is monotone decreasing in $u$. Hence, $0 \leq \varepsilon_0 \leq D(f_1, f_0)$ and $0 \leq \varepsilon_1 \leq D(f_0, f_1)$.

Proof: For $j = 0$, it follows that
\[ \varepsilon(u) = -\ln (k(u)) + \frac{u}{k(u)} \int_{\mathbb{R}} w(y; u) \ln l(y) f_0(y) dy \]

After manipulation, the first derivative of $\varepsilon(u)$ is
\[ \frac{\partial \varepsilon(u)}{\partial u} = \frac{u}{k(u)^2} \left[ k(u) \int_{\mathbb{R}} f_0(y) (l(y) u \ln l(y))^2 dy + \frac{-k(u)}{\partial u} \int_{\mathbb{R}} (l(y) u f_0(y) \ln l(y) dy \right]. \]
Inserting $\partial^2 \varepsilon(u)/\partial u$ and rearranging the terms yields
\[
\frac{k(u)^2 \varepsilon(u)}{u^2 \partial u} = \int_{\mathbb{R}} \left( l(y) f_0(y) \ln(l(y)) \right)^2 dy
- \int_{\mathbb{R}} l(y)^u f_0(y) \ln(l(y)) dy \int_{\mathbb{R}} l(y)^u f_0(y) \ln(l(y)) dy
+ \int_{\mathbb{R}} w(y; u) dy \int_{\mathbb{R}} w(y; u) \ln(l(y))^2 \, dy
- \left( \int_{\mathbb{R}} w(y; u) \ln(l(y)) dy \right)^2
(52)
\]
By Hölder’s inequality, $w(y; u)$ is integrable over $\mathbb{R}$. Consider the weighted $L^2$ space, $L^2_w(\mathbb{R})$ equipped with the inner product
\[
(g, h)_w := \int_{\mathbb{R}} g(y) h(y) w(y) \, dy
(53)
\]
and the resulting norm $\|g\|_w = \sqrt{(g, g)_w}$. By definition, $g$ is in $L^2_w$ if $g^2 w$ is integrable over $\mathbb{R}$. Let $g(y) = \ln(l(y))$. Dividing (52) by $(\int_{\mathbb{R}} w(y) \, dy)^2$ reads
\[
\frac{k(u)^2 \varepsilon(u)}{u^2 \left( \int_{\mathbb{R}} w(y) \, dy \right)^2 \partial u} = \left\| g \right\|_w^2 - (g, 1)_w^2
= \left\| g \right\|_w^2 \left\| 1 \right\|_w^2 - (g, 1)_w^2 > 0
(54)
\]
The inequality follows from the Cauchy-Schwarz inequality for the inner product space $(g, h)_w$ and it is strict since $g$ and $1$ are linearly independent. What remains to be shown is that $g$ belongs to $L^2_w$, i.e., $\int_{\mathbb{R}} g(y)^2 w(y) \, dy < \infty$. If $g$ is bounded, the claim is obvious. If not, then, either $\lim_{|y| \to \infty} l(y) = \infty$ or $\lim_{|y| \to \infty} l(y) = 0$. Assume $\lim_{|y| \to \infty} l(y) = \infty$ and write
\[
\ln(l(y))^2 w(y) = (\ln(l(y))^2 l(y)^u f_0(y)
= \ln(l(y))^2 \frac{l(y)^u}{l(y)^{1/u}} f_0(y)
= \ln(l(y))^2 \frac{f_1(y)^{1/u}}{l(y)^{1/u}} f_0(y)^{1/u}
(55)
\]
By Hölder the function $f_1(y)^{1/u} f_0(y)^{1/u}$ is integrable and since
\[
\lim_{|y| \to \infty} \ln(l(y))^2 = 0,
(56)
\]
g(y)^2 w(y) is integrable over $[0, \infty)$ by comparison with $f_1(y)^{1/u} f_0(y)^{1/u}$. If $\lim_{|y| \to \infty} l(y) > 0$, then $g$ is bounded on $(-\infty, 0]$ and integrability over $(-\infty, 0]$ follows. If $\lim_{|y| \to \infty} l(y) = 0$, then as
\[
\ln(l(y))^2 w(y) = (\ln(l(y))^2 l(y)^u f_0(y),
(57)
\]
we have
\[
\lim_{y \to \infty} \ln(l(y))^2 l(y)^u = 0,
(58)
\]
and integrability over $(-\infty, 0)$ follows by comparison with $f_0$. In a similar way, $g(y)^2 w(y)$ is integrable over $\mathbb{R}$ if $\lim_{y \to -\infty} l(y) = \infty$ or $\lim_{y \to \infty} l(y) = 0$. This completes the proof that $\partial^2 \varepsilon(u)/\partial u > 0$ and hence $\varepsilon_0 \leq \varepsilon_0(1) = D(f_1, f_0)$. For $\varepsilon_1$, let $u = 1 - u$, $f_1 := f_0$ and $f_0 := f_1$. This gives $\varepsilon_1(u') = \varepsilon_0(u')$, which implies that $\varepsilon_1(u')$ is increasing, therefore $\varepsilon_1(u)$ is decreasing. Note that for $\varepsilon_1$, with the substitutions of the densities, $l$ becomes decreasing, however $g$ still belongs to $L^2_w$ and the proof is complete. ■

Prop. (III.2) implies that (50) has a unique solution for all $\varepsilon_j \leq D(f_{1-j}, f_j)$, $j \in \{0, 1\}$. In particular, given a certain choice of $\varepsilon_j$, the solution of Eq. (50) leads to $0 \leq u^* \leq 1$. The corresponding maximum $\varepsilon_{1-j}(u^*)$ is therefore obtained by $\varepsilon_{1-j}(u^*)$. From (50), it also follows that
\[
\varepsilon_0(u) - \varepsilon_1(u) = \frac{1}{k(u)} \int_{\mathbb{R}} l(y)^u f_0(y) \ln(l(y)) \, dy,
(59)
\]
which is bounded as $-D(f_0, f_1) \leq \varepsilon_0(u) - \varepsilon_1(u) \leq D(f_1, f_0)$ due to monotonicity. When $\varepsilon = \varepsilon_0(u) = \varepsilon_1(u)$, this reduces to
\[
\varepsilon = \sup_{0 \leq u \leq 1} - \ln \int_{\mathbb{R}} f_1(y)^u f_0(y)^{1-u} \, dy
(60)
\]
which is the Chernoff distance and if additionally $f_0(y) = f_1(-y)^2 y$, it further reduces to
\[
\varepsilon = - \ln \int_{\mathbb{R}} \sqrt{f_0(y) f_1(y)} \, dy,
(61)
\]
which is the Bhattacharyya distance between the nominal densities.

C. Limiting robustness parameters for the $(h)$-test

**Proposition III.3.** The maximum achievable pair of $(\varepsilon_0, \varepsilon_1)$ with respect to the $e$-contamination model are obtained by
\[
(1 - \varepsilon_0) |P_0[l \leq 1/b] - b P_1[l \leq 1/b]| = \varepsilon_1
(62)
\]
where $b = (1 - \varepsilon_1)/(1 - \varepsilon_0)$.

**Proof:** By Huber [5], it is known that $h_1(c_1) = P_1[p_1/p_0 > c_1] + c_1 P_0[p_1/p_0 \leq c_1]$ is an increasing function of $c_1$ and $h_2(c_0) = P_0[p_1/p_0 < c_0] + (1/c_0) P_1[p_1/p_0 \geq c_0]$, in a similar manner, is a decreasing function of $c_0$. This implies that $\varepsilon_0$ and $\varepsilon_1$ are maximized when $c_1$ is maximized and $c_0$ is minimized. The maximum of $c_1$ is equal to the minimum of $c_0$ such that the hypotheses do not overlap. As a result for $c = c_1 = c_0$, it follows that $\hat{l}(y) = bc$ for all $y \in \mathbb{R}$. Since no density is greater than any other for all $y \in \mathbb{R}$, the conclusion is that $c = 1/b$. Rewriting the equations, $h_1(c := 1/b) = 1/(1 - \varepsilon_1)$ or equivalently $h_2(c := 1/b) = 1/(1 - \varepsilon_0)$, completes the proof. ■

Let $u_0 = \text{ess inf}[u], u_1 = \text{ess sup}[u], l_1 = \text{ess inf}[l]$, and let $k = 1 - \varepsilon_0$ be known and $u = 1/(1 - \varepsilon_1)$ to be determined. With these substitutions (62) can be written as $f(u) = ku P_0[l \leq ku] - P_1[l \leq ku] - u + 1$.

**Lemma III.4.** The function $f$ is continuous, $f(u) = 1 - u$ for $0 \leq k \leq u_0$, strictly decreasing for $ku < u_1$, tends to $-\infty$ for $k < 1$ and tends to $0$ for $k = 1$ as $u \uparrow \infty$.

**Proof:**
\[
f(u) = ku P_0[l \leq ku] - P_1[l \leq ku] - u + 1
= \int_{\{l \leq ku\}} (ku - l) p_0 d\mu - u + 1.
(63)
Thus,
\[ f(u + \Delta) - f(u) = \int_{\{ku < t \leq k(u + \Delta)\}} (ku + k\Delta - l) p_0 d\mu \\
+ \Delta \left( k \int_{\{t \leq ku\}} p_0 d\mu - 1 \right). \]  
(64)

It then follows that
\[ f(u + \Delta) - f(u) \leq k\Delta P_0[|ku - \Delta| \leq k(u + \Delta)] \\
+ k\Delta P_0[|l - ku| - \Delta] \\
= \Delta (kP_0[|l - ku| - \Delta] - 1) \leq 0 \]  
(65)
for any positive \( \Delta \) and for all \( u \geq 0 \). Since \( f(u + \Delta) - f(u) \geq \Delta (kP_0[|l - ku| - 1] - \Delta) \), the conclusion is that \( 0 \geq f(u + \Delta) - f(u) \geq \Delta \) from where continuity and monotonicity follow. For \( ku \leq k(u + \Delta) < u_1 \), it also follows that \( f(u + \Delta) - f(u) < 0 \), hence \( f(u) \) is strictly decreasing, tends to \(-\infty\) for \( k < 1 \) and tends to 0 for \( k = 1 \) as \( u \uparrow \infty \).

Lemma [III.4] implies that \( \epsilon_1 \) can be determined uniquely for all \( 0 \leq \epsilon_0 < 1 \). Moreover, Lemma [III.4] extends to the case when \( \epsilon_1 \) is known and \( \epsilon_0 \) is variable due to the duality of the parameters, \( \epsilon_0 \) and \( \epsilon_1 \).

IV. FIXED SAMPLE SIZE TESTS

The robust version of the likelihood ratio test with respect to the uncertainty model (1) can be generalized to \( n \) independent samples, i.e.,
\[ \hat{I}(y) = \prod_{i=1}^{n} \hat{I}(y_i) \overset{\mathcal{H}_0}{\gtrless} \gamma, \]  
(66)
which is equivalent to the nominal likelihood ratio test
\[ \prod_{i=1}^{n} I(y_i) \overset{\mathcal{H}_0}{\gtrless} \gamma, \]  
(66)
and similarly in the logarithmic scale
\[ \sum_{i=1}^{n} \ln \left( I(y_i) / (l_i / u_i) \right) \overset{\mathcal{H}_0}{\gtrless} n \ln l_i, \]  
(67)
for \( \gamma = 1 \). Given the upper and lower thresholds, \( l_i \) and \( u_i \), if \( \sum_{i=1}^{n} \delta_i(y_i) = 0 \), the original threshold of the nominal test is moved from 0 to \( n \ln l_i \), increasing the false alarm probability. Similarly, if \( \sum_{i=1}^{n} \delta_i(y_i) = 1 \), the original threshold of the nominal test is moved to \( n \ln u_i \), which increases the miss detection probability. Let \( \mathbf{y} = [y_1, \ldots, y_n] \) be the observation vector and \( b = 1 \). Assume that there are \( n_1 \) (and \( n_2 \)) observations in \( y \) whose likelihood ratios are clipped to \( (c_l) \) (and \( c_u \)), respectively. Then, Huber’s clipped likelihood ratio test can be represented in the log domain as
\[ \sum_{i=1}^{n-n_1-n_2} \ln \left( I(y_i) \right) \overset{\mathcal{H}_0}{\gtrless} -(n_1 \ln c_u + n_2 \ln c_l). \]  
(69)
Eventually, the robust test based on the composite model can be given by
\[ \sum_{i=1}^{n-n_1-n_2} \ln \left( I(y_i) I_1 - \delta_i(y) I_l / \delta_i(y) I_u \right) \overset{\mathcal{H}_0}{\gtrless} -(n_1 \ln c_u + n_2 \ln c_l). \]  
(70)

where \( n_1 \) and \( n_2 \) are now due to clipping of the likelihood ratio given by (15). The composite test combines the robustness properties of both the clipped likelihood ratio test as well as the robust test for modeling errors. Single sample robust tests are extended to multiple samples through multiplication of the likelihood ratios due to the independency of every measurable set of observations. Unlike Huber’s robust test, there is no stochastic ordering for the LFDs of modeling errors. Hence, the composite model can be expected to be robust, but minimax robustness is not guaranteed for \( n > 1 \).

A. Asymptotic performance analysis

Large deviations theory can be used to analyze the asymptotic performance of the robust tests. Consider the following theorem by Cramér [19]:

**Theorem IV.1** (Cramér). Let \( (Y_i)_{i=1}^{\infty} \) be a sequence of i.i.d. random variables, \( S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \) be their average sum and \( \widehat{M}_Y(u) := E[e^{uY_1}] < \infty \) be the moment generating function of the r.v. \( Y_1 \). Then, for all \( t > E[Y_1] \)
\[ \lim_{n \to \infty} \frac{1}{n} \ln P(S_n \geq t) = -I(t) \]  
(71)
where the rate function \( I \) is defined by
\[ I(t) := \sup_u \{ tu - \ln M_Y(u) \}, \]  
(72)
which is the Legendre transform of the log moment generating function.

**Remark IV.1.** Theorem [IV.1] implies
\[ \lim_{n \to \infty} \frac{1}{n} \ln P(S_n < t) = -I(t) \]  
(73)
for all \( t < E[Y_1] \). To see this, take \( X_i = -Y_i \) and consider
\[ P \left( \frac{1}{n} \sum_{i=1}^{n} X_i > -t \right). \]  
(74)

Applying Cramér’s theorem to the r.v. \( X_1 \) and the threshold \(-t\), it follows that \( M_{X_1}(u) = M_{X_1}(-u) \) and
\[ I(t) := \sup_u \{ tu - \ln M_{X_1}(u) \} = \sup_u \{ -tu - \ln M_{X_1}(-u) \} \]  
(75)
Let \( S_n^* := \frac{1}{n} \sum_{i=1}^{n} \ln \hat{I}(Y_i) \) with \( \hat{I}(Y_i) = \hat{q}_1(Y_i) / \hat{q}_0(Y_i) \) for \( Y_i \sim Q_0 \) under \( \mathcal{H}_0 \) and for \( Y_i \sim Q_1 \) under \( \mathcal{H}_1 \) for all \( i \in \{0, \ldots, n\} \). Furthermore, let the first and second type of error probabilities defined to be \( P_{E_1}(t) = P(S_n^* \geq t) \) and \( P_{E_2}(t) = P(S_n^* < t) \). Then, for all \( E_{Q_0} [\hat{I}(Y_1)] < t < E_{Q_1} [\hat{I}(Y_1)] \) from Theorem [IV.1] and Remark [IV.1]
\[ \lim_{n \to \infty} \frac{1}{n} \ln P_{E_1}(t) = -I_j(t) \quad j = 0, 1, \]  
(76)
where
\[ I_j(t) := \sup_u \left( tu - \ln M_{Y_1}^j(u) \right) \quad j = 0, 1, \]  
(77)
with
\[ M_{Y_1}^j(u) = \int_{\mathbb{R}} \hat{I}(y) q_j(y) dy \quad j = 0, 1. \]  
(78)
Remark IV.2. Interestingly, if \( \hat{q}_1 = q_1 \) and \( \hat{q}_0 = q_0 \), the parametric curve \((\varepsilon_0(u), \varepsilon_1(u))\) for \(0 \leq u \leq 1\), with \( f_0 := q_0 \) and \( f_1 := q_1 \) implies \((I_0(t), I_1(t))\) for all \(E_{Q_0}[\ln(Y_1)] < t < E_{Q_1}[\ln(Y_1)]\). To prove this claim, observe that in this case we have \( M_1^0(u) = M_1^0(u+1) \). Applying this result to (77), taking the derivative of \(tu - \ln M_1^0(u)\) with respect to \(u\), and rewriting \( I_j\) in terms of maximizing \(u\) gives (50) with the aforementioned substitutions. Since the mapping from \(0 \leq u \leq 1\) to \(E_{Q_0}[\ln(Y_1)] < t < E_{Q_1}[\ln(Y_1)]\) is bijective, as the derivative of a convex function \(\ln M_1^0(u)\) is increasing, the proof is complete.

B. Limiting tests

1) Limiting (m)-test: The limiting case, \( \lim_{t \to \infty} I_t \) and \( \lim_{n \to \sup} l_n\), is of particular interest. For a single sample, the test becomes a pure randomized test having a success probability \(\delta\) which increases with \(l\) (14). For \(n\) independent samples, assume \(l_i := 1/n \) and consider the normalization \(\ln l(y) = (\ln l_u - \ln l(y))/(\ln l_u - \ln l_i)\). Then, as \(l_i \downarrow 0\) and \(l_u \uparrow \infty\), the test statistic \(\ln l_n(y) = \sum_{i=1}^n \ln l_i(y_i)\) tends to \(\sum_{i=1}^n \delta(y_i)\), which is the soft version of the sign test.

2) Limiting (h)-test: The limiting test for Huber’s clipped likelihood ratio test is known to be the sign test [5].

3) Limiting (a)-test: The limiting asymptotically robust test is again a likelihood ratio test with the threshold determined by \(u + v \to 1\) in [35].

V. ROBUST SEQUENTIAL PROBABILITY RATIO TEST

Sequential probability ratio tests (SPRTs) can be preferable over fixed sample size tests due to their strong optimality properties [2]. Let \(S_n = \sum_{i=1}^n Y_i\). Then, for given target error probabilities of the first and second kind, \(\alpha\) and \(\beta\) respectively, by Wald [29], there exist an upper threshold \(t_u > 1\) and a lower threshold \(0 < t_l < 1\) such that SPRT continues taking another sample if \(t_l < S_n < t_u\), terminates and decides for \(\mathcal{H}_0\) if \(S_n < t_l\) and decides for the alternative hypothesis \(\mathcal{H}_1\) if \(S_n > t_u\), for the first time \(N = \min\{n : S_n > t_u \text{ or } S_n < t_l\}\). Furthermore, let the binary r.v. \(v\) denote the decision of the sequential test, i.e. \(v = 1\) to decide for \(\mathcal{H}_1\) and \(v = 0\) for \(\mathcal{H}_0\). Similar to the fixed sample size test, a robust version of the sequential test can be defined in terms of the nominal likelihood ratios and modified thresholds

\[
m(y_i) \ln t_l < \sum_{i=1}^{n} \ln l(y_i) < m(y_i) \ln t_u
\]

with

\[
m(y_i) = \ln l_i \left( n - \sum_{i=1}^{n} \delta(y_i) \right) + \ln l_u \sum_{i=1}^{n} \delta(y_i),
\]

for the (m)-test. Extensions to the (h)-test as well as to the (c)-test for the function \(m\) follow in a straightforward manner from (59) and (70). However, it can be observed that all three robust tests are still some subsets of a possible design which considers two possibly different functions \(m_0 : \Omega \to \mathbb{R}\) and \(m_1 : \Omega \to \mathbb{R}\) as multicplicand to the lower and upper thresholds. Hence, it can be concluded that a general design of a robust sequential test is a design of two (random) functions \(m_0\) and \(m_1\) such that both the expected number of samples, \(E[N]\), as well as the error probabilities of the first and second kind \((\alpha, \beta)\) are bounded from above for all probability measures in the vicinity of the nominal distributions defined by a neighborhood of uncertainty. In the following, the robust tests that have already been designed or introduced are analyzed for the sequential test. Throughout design or analysis of a robust sequential test can be found for example in [21], where the probability distributions are assumed to be discrete with finite set of values, or in [22], where Huber’s test is rigorously shown to be asymptotically robust. Let \(S_n = \sum_{i=1}^n \ln l(Y_i)\) be the test statistic where \(Y_1, Y_2, \ldots, Y_n\) are again i.i.d. and \(\ln l(Y_i)\) follows a probability distribution \(Q_j\), which accepts a continuous density function \(q_j\), when the true hypothesis is \(\mathcal{H}_j, j \in \{0, 1\}\). Let furthermore

\[
h_{j,n}(y) = \frac{\partial}{\partial y} P_j[S_n \leq y, S_1, \ldots, S_{n-1} \in (ln t_l, ln t_u)] \tag{81}
\]

be the density function of \(S_n\) under \(\mathcal{H}_j, j \in \{0, 1\}\) when all \(S_k, k < n\) are in \((ln t_l, ln t_u)\). Hence, the distribution of \(N\) can be calculated recursively by

\[
P_j[N = n] = \int_{(-\infty, ln t_l) \cup (ln t_u, \infty)} h_{j,n}(y)\,dy
\]

\[
h_{j,n}(y) = \int_{ln t_l}^{ln t_u} h_{j,n-1}(\omega)q_j(y - \omega)\,d\omega
\]

with the initial condition \(h_{j,1} = q_j, j \in \{0, 1\}\). Accordingly, it follows that

\[
E_j[N] = \sum_{n=1}^{\infty} n P_j[N = n]. \tag{83}
\]

Slightly modifying \(P_j[N = n]\) by imposing the constraint that the test will terminate either with the rejection or acceptance \(\mathcal{H}_0\),

\[
P_0[N = n|v = 1] = \int_{(ln t_u, \infty)} h_{0,n}(y)\,dy
\]

\[
P_1[N = n|v = 0] = \int_{(-\infty, ln t_l)} h_{1,n}(y)\,dy
\]

we get

\[
\alpha = \sum_{n=1}^{\infty} P_0[N = n|v = 1], \beta = \sum_{n=1}^{\infty} P_1[N = n|v = 0]. \tag{85}
\]

Herein, \(\alpha, \beta,\) and \(E_j[N]\) are all implicit functions of \((t_l, t_u)\), and \(Q_j, j \in \{0, 1\}\). When the notations are made explicit, a minimax robust sequential test must satisfy

\[
\alpha_{Q_0}[t_l, t_u] \geq \alpha_{Q_0}[t_l, t_u]
\]

\[
\beta_{Q_1}[t_l, t_u] \geq \beta_{Q_1}[t_l, t_u]
\]

and

\[
E_{Q_0}[N|t_l, t_u] \geq E_{Q_0}[N|t_l, t_u]
\]

\[
E_{Q_1}[N|t_l, t_u] \geq E_{Q_1}[N|t_l, t_u]
\]

\[
E_{Q_0}[N|t_l, t_u] \geq E_{Q_0}[N|t_l, t_u]
\]

\[
E_{Q_1}[N|t_l, t_u] \geq E_{Q_1}[N|t_l, t_u]
\]
for all \((Q_0, Q_1) \in Q_0 \times Q_1\) and for all \((t_l, t_u)\).

The sequential \((m)\)-test does not satisfy \((86)\) and \((87)\) even asymptotically, i.e., when \(t_l \to 0\) and \(t_u \to \infty\), or equivalently \(Q_0 \to Q_1\) or \(Q_1 \to Q_0\). This is due to the lack of stochastic ordering between \(Q_0\) and \(Q_0\) and between \(Q_1\) and \(Q_1\).

Similarly, the sequential \((a)\)-test \((33)\) does not satisfy \((86)\) (asymptotically) either. Again, asymptotically, the behavior of the cumulative sums are determined by their non-random drift, i.e., \(S_n \sim NE_Q[\ln i(Y)]\) or \(S_n \sim NE_Q[\ln i(Y)]\) and Wald's approximations become exact, i.e., \(E[S_n] \approx \ln t_l\) under \(H_0\) and \(E[S_n] \approx \ln t_u\) under \(H_1\). Combining both conditions, it follows that

\[
E_{Q_0}[N] \sim \frac{\ln t_l}{E_{Q_0}[\ln i(Y)]}, \quad E_{Q_1}[N] \sim \frac{\ln t_u}{E_{Q_1}[\ln i(Y)]}.
\]

(88)

From \((36)\), it is known that \((33)\) maximizes the right hand sides of \((88)\). Therefore, the sequential \((a)\)-test satisfies \((87)\) asymptotically.

For the sequential \((h)\)-test, it is known that \((86)\) and \((87)\) are satisfied asymptotically \([5]\). Additionally in \([4]\), a counterexample is given, which shows that \((87)\) does not hold in general, i.e., for all \((t_l, t_u)\). In the following, it is shown that the sequential \((h)\)-test satisfies \((86)\) for all \((t_l, t_u)\).

**Theorem V.1** (Coupling). Let \((X, Y)\) be a pair of random variables on \((\Omega, \mathscr{A}, P)\) with \(X \succ_T Y\). On the same probability space there exist another pair of random variables \((X', Y')\) such that \(X = X'\) in distribution, \(Y = Y'\) in distribution and \(X' \geq Y'\) almost surely.

**Proof:** Take \(X' = X\) and \(Y' = G^{-1}(F(X))\). Then, \(X' = X\) in distribution, \(P(G^{-1}(F(X)) \leq x) = P(F(X) \leq G(x)) = P[X \leq F^{-1}(G(x))] = F[F^{-1}(G(x))] = G(x) =: P[Y \leq x]\), so \(Y = Y\) in distribution and since \(P[Y' \geq X'] = P[G^{-1}(F(X)) \geq X] = P[F(X) \geq G(X)] = 1\), \(X' \geq Y'\) almost surely.

**Proposition V.2.** Let \(X_1\) and \(Y_1\) be two continuous random variables on \(\mathbb{R}\) having distribution functions \(F\) and \(G\), respectively, and satisfying \(G(y) \geq F(y)\) for all \(y\). Furthermore, let \(S_n^X = \sum_{i=1}^n X_i\), \(S_n^Y = \sum_{i=1}^n Y_i\), \(A > 0\), and \(B < 0\). Denote \(\tau_A = \inf\{n \geq 0 : S_n \geq A\}\) and \(\tau_B = \inf\{n \geq 0 : S_n \leq B\}\) the hitting/stopping times of \(S_n\) at the upper and lower thresholds respectively, then,

\[
P_{S_n^X}[\tau_A > \tau_B] \geq P_{S_n^Y}[\tau_A > \tau_B].
\]

(89)

**Proof:** For a well defined comparison, exclude the cases \(X \neq 0\) and \(Y \neq 0\) s.t. at least \(\tau_A < \infty\) or \(\tau_B < \infty\) almost surely and \(\tau_A > \tau_B\) is well defined. The argument \(G(y) \geq F(y)\) for all \(y\) implies \(X \succ_T Y\) and from Prop. V.1 there exists \((X', Y')\) such that \(X' = X\), \(Y' = Y\) in distribution and \(X' > Y'\) almost surely (a.s.). Consider the sequence of i.i.d. random variables \((X'_n, Y'_n)_{n \geq 1}\) s.t. \((X'_n, Y'_n) = (X', Y')\) in distribution. Then, \((X_n)_{n \geq 1} = (X'_n)_{n \geq 1}\) and \((Y_n)_{n \geq 1} = (Y'_n)_{n \geq 1}\) in distribution. Defining \(S_n^X = \sum_{i=1}^n X_i\) and \(S_n^Y = \sum_{i=1}^n Y_i\), we also have \(S_n^X = S_n^X\) and \(S_n^Y = S_n^Y\) in distribution. Since \(X' > Y'\) a.s. and accordingly \(X_i > Y_i\) a.s. for all \(i\), \(S_n^X \geq S_n^Y\) a.s. Let \(\tau'_{X} = \inf\{n \geq 0 : S_n^X \geq A\}\) and define \(\tau'_{X'}\), \(\tau'_{Y'}\) and \(\tau'_{Y'}\) in the same way. Then \(S_n^{X'} \geq A\) implies \(S_n^X \geq A\) for all \(n\), so \(\tau'_{X} \leq \tau'_{X'}\) and in the same way \(\tau'_{Y} \geq \tau'_{Y'}\). Hence,

\[
P(S_n^X[\tau_A < \tau_B] = P(\tau'_{X} < \tau'_{X'}) \geq P(\tau'_{X} < \tau'_{Y'}) = P(S_n^Y[\tau_A < \tau_B])
\]

(90)

Let \(X \sim \hat{Q}_0\) and \(Y \sim Q_0\), likewise \(X \sim Q_1\) and \(Y \sim \hat{Q}_1\) with \(A = \ln t_u\) and \(B = \ln t_l\). Then, it is easy to see that \((90)\) is equivalent to \((86)\) for any pair \((t_l, t_u)\). This result includes not only the (h)-test, but also all tests in \([3]\), \([9]\).

For the expected number of samples, the requirement is

\[
E[\min\{\tau'_{X'}, \tau'_{Y'}\}] \geq E[\min\{\tau'_{X'}, \tau'_{Y'}\}].
\]

(91)

This inequality does not have to hold in general. Intuitively, however, it is expected that it holds for the majority of the cases, especially when \(t_l\) is small enough and \(t_u\) is large enough.

**VI. Robust estimation**

The composite uncertainty model given in equation \((39)\) extends to robust estimation problems. Let \(f_0\) be a nominal probability density function corresponding to the distribution function \(F\) with parameters \(\theta = [\theta_1, \theta_2, \ldots, \theta_N]\). In a general estimation framework, some parameters, possibly a sub-vector of \(\theta\) can be estimated well whereas some other parameters might not be, possibly due to a fast change of the parameters with time or due to the random nature of the parameters whose distributions are unknown. It is also possible that the known parameters might deviate slightly from the true values depending on the nature of the application or without regarding the parametric model, the shape of the distribution might be slightly different than expected, e.g. when there is lack of data but the CLT is assumed. In such cases, we have modeling errors that go unmodeled in addition to the outliers caused by some unexpected events. Therefore, it is desirable to design robust estimators which are not only able to deal with outliers but also with modeling errors, as given by \((39)\).

To account for the composite model, let \(T_n(F)\) be a functional \(T_n : Y_n \to \mathbb{R}\) of \(\mathbb{R}^n\)-valued random variable \(Y_n\) with i.i.d. replicas following a certain distribution \(F\), i.e., for \(Y_n = [Y_1, Y_2, \ldots, Y_n]\) each pair of r.v.s \((Y_i, Y_k)\) with \(i \neq k\) are i.i.d., having a distribution function \(F\). Then, it is desirable that \(\lim_{n \to \infty} T_n(F) = \theta\) for some parameter \(\theta\) when \(F\) is the nominal distribution. Let \(F_{T_n}\) and \(Q_{T_n}\) be the distribution functions of \(T_n\) when \(F\) and \(Q\) are the distribution functions of \(Y_1\), respectively. Then, it is also expected that for every \(\epsilon > 0\), there exist \(\delta > 0\) and an \(n_0 > 0\) such that for all \(n > n_0\) and \(Q \in \mathcal{F}\), \(D(F_{T_n}, Q_{T_n}) < \epsilon\) whenever \(D(F, Q) < \delta\) for some metric \(D\). This is a straightforward extension of Hampel’s equicoherence theorem of robustness for the composite uncertainty model. Accordingly, the influence function can be modified as

\[
IF(y, T) = \lim_{\epsilon \to 0} \sup_{G \in \mathcal{G}} \frac{T((1-\epsilon)G + \epsilon\delta_x) - T(F)}{\epsilon}
\]

(92)
to account for the modeling errors in addition to the outliers. Similarly, the maximum bias as being another important metric to measure the robustness of an estimator can be obtained as

$$b(\epsilon) = \sup_{Q,G} |T(Q) - T(G)| = \sup_{G,H} |T((1 - \epsilon)G + \epsilon H) - T(G)|.$$  

(93)

VII. SIMULATIONS

In this section, simulations are performed in order to visualize and validate the theoretical findings. Observations are assumed to be real valued. The formulations are general, therefore, the observation space can be any discrete, continuous, finite or infinite set, with slight modifications for the discrete case. It can also be extended to the multidimensional case, but for large $n$, Monte-Carlo simulations may be required in order to solve the non-linear equations, c.f., [10].

In the first simulation, the composite uncertainty model [39] with mean and variance shifted nominal distributions, $F_0 \sim \mathcal{N}(-1.1)$ and $F_1 \sim \mathcal{N}(1.2)$, and the uncertainty parameters \( \{\varepsilon_0 = 0.15, \varepsilon_1 = 0.05, \epsilon_0 = 0.02, \epsilon_1 = 0.02\} \) is considered. Note that for this choice of nominal distributions, neither $\partial F_1 / \partial F_0$ is monotone nor they are symmetric with respect to any point on their domain or codomain. In addition to this, $\varepsilon_0 \neq \varepsilon_1$ is chosen so that the given example is general enough for the solution of Equations (16) and (17). Regarding the $\varepsilon$–contamination part of the composite model, $\varepsilon_0 = \varepsilon_1$ is chosen to be consistent with $\rho = 1$ for the uncertainty model based on relative entropy. Accordingly, in Fig. 1 the log-likelihood ratios of the nominal distributions, the least favorable densities $\hat{g}_0, \hat{g}_1$ when $\{\varepsilon_0 = 0.15, \varepsilon_1 = 0.05, \epsilon_0 = \epsilon_1 = 0\}$ and the least favorable densities $\hat{q}_0, \hat{q}_1$ when $\{\varepsilon_0 = 0.15, \varepsilon_1 = 0.05, \epsilon_0 = 0.02, \epsilon_1 = 0.02\}$ are shown.

In the second simulation, the mean shifted Gaussian distributions $F_0 \sim \mathcal{N}(-1.1)$ and $F_1 \sim \mathcal{N}(1.1)$ are considered when the closed balls are formed with respect to the symmetrized $\chi^2$ distance with $\varepsilon_0 = \varepsilon_1 = 0.08$ and a relative entropy distance $D$ with $\varepsilon_0 = \varepsilon_1 \approx 0.0087$. The parameters are chosen such, such that the LFDs resulting from both distances have equal relative entropy relative to the nominal density functions. Figure 2 illustrates $l/l$, the ratio of the likelihood ratios. It can be seen that there is a significant difference when the $\chi^2$ distance is considered instead of the KL-divergence. While this ratio tends to 1 as $\delta \to 0$ and $\delta \to 1$ for the symmetrized $\chi^2$ distance, meaning that the tails of the density functions are preserved, it is a constant $l_i < 1$ when $\delta = 0$ and another constant $l_i > 1$ when $\delta = 1$ for the KL-divergence.

In the third simulation, again the same mean shifted Gaussian distributions are considered. Of interest is the curvature of the maximum robustness parameters for the (h)-test (62) versus the (m)-test (50). Figure 3 illustrates the outcome of this simulation.

In the fourth simulation, asymptotic decrease rates, $I_0$ and $I_1$ (77), of the type I and type II error probabilities are considered. The log-likelihood ratio test is built based on LFDs of the composite model $l(Y) = \hat{q}_1(Y) / \hat{q}_0(Y)$ with parameters $\varepsilon_0 = \varepsilon_1 = 0.01$ and $\epsilon_0 = \epsilon_1 = 0.01$. The r.v.s $Y_0, Y_1, \ldots, Y_n$, which are consistent with the observations.
$y_0, y_1, \ldots, y_n$, are i.i.d. The simulation is performed for six different distributions of $Y_1$ for each hypothesis. Under $\mathcal{H}_j$, $Y_1$ is distributed as one of the following distributions: the nominal distribution $F_1$ denoted by (n), LFD $Q_j$ with parameters $\epsilon_0 = \epsilon_1 = 0.01$ denoted by (m), LFD $Q_i$ with parameters $\epsilon_0 = \epsilon_1 = 0.01$ denoted by (h), LFD $\hat{G}_j$ of the asymptotically robust test with $\epsilon_0 = \epsilon_1 = 0.01$ denoted by (a), LFD of the composite model $\hat{Q}_j$ with parameters $\epsilon_0 = \epsilon_1 = 0.01$ and $\epsilon_0 = \epsilon_1 = 0.01$ denoted by (c), $j \in \{0, 1\}$. For comparison reasons, the sixth LFD is introduced with respect to the composite uncertainty set. The LFD of the asymptotically robust test $\hat{G}_j$ for $\epsilon_0 = \epsilon_1 = 0.01$ are first obtained. Then, $\hat{Q}_j$ with $\epsilon_0 = \epsilon_1 = 0.01$ is determined when $\hat{G}_j$ is the nominal distribution, $j \in \{0, 1\}$. This test is denoted by (c*). Figure 5 and Fig. 3 illustrate $I_0$ and $I_1$ when $Y_1$ follows various distributions, as described above. The notation [8] indicates that the robust test is performed by the LFDs of the (a)-test and the observations follow the LFD of (b)-test. In general, the composite test is not claimed to be asymptotically minimax robust since the LFDs of the (m)-test are not asymptotically robust. However, for this example, the (c)-test asymptotically does not degrade its performance for all observation models, when $t$ is small enough in its allowable limits. This test corresponds to the type I Neyman-Pearson test, cf. [2].

In the fifth simulation, a single sample (m)-test [14] is considered, when the nominal distributions are mean shifted and mean and variance shifted Gaussian distributions as defined before. Robustness parameters are chosen to be equal ($\epsilon = \epsilon_0 = \epsilon_1$). For this choice, from [50], it follows that $\epsilon \in [0, 0.5]$ for the mean shifted Gaussian distributions and $\epsilon \in [0, 0.338]$ for the mean and variance shifted Gaussian distributions s.t. the LFDs do not fully overlap. For all possible choices of $\epsilon$, the performance of this robust test was calculated when the observations are due to LFDs of the (m)-test [13] and the LFDs of the (a)-test [33], which are determined for the same $\epsilon$ of the robust test. The rationale behind this simulation is to test the minimax property defined by [9] and [10]. The choice of the (a)-test as a competitor to the (m)-test is not arbitrary. First, the LFDs of both tests lie on the boundary of the closed ball and second, the (a)-test is claimed to be asymptotically robust for large enough $n$ [6]. Figure 7 illustrates the outcome of this simulation for the mean shifted Gaussian distributions. Due to the symmetry of the nominal distributions and the equal choice of the robustness parameters, we have $P_{\text{F}} = P_{\text{F}m} = P_{\text{F}a}$. It can be seen that the robust test doesn’t degrade its performance as expected. Similarly, in Fig. 8 the result of the same simulation for the mean and variance shifted Gaussian distributions is given. Since the nominal distributions are not symmetric, the error probabilities ($P_{\text{F}m}^n$ and $P_{\text{F}a}^n$) are unequal. More interestingly, as illustrated in Fig. 9 the false alarm probability first increases with $\epsilon$ and then starts decreasing. In all cases, it can be seen that (9) and (10) are valid.

The last part of the simulations is related to the robustness of the sequential probability ratio test based on the likelihood

![Figure 4](image-url) Maxima. A new robustness parameter with respect to the (h)-test and the (m)-test when the nominal distributions are $F_0 \sim \mathcal{N}(-1, 1)$ and $F_1 \sim \mathcal{N}(1, 1)$.

![Figure 5](image-url) Asymptotic decrease $I_0$ of the composite test when the observations follow the nominal distributions $F_0 \sim \mathcal{N}(-1, 1)$ and $F_1 \sim \mathcal{N}(1, 2)$, LFDs of the (m)-test, LFDs of the (a)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, LFDs of the (h)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, LFDs of the (c)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, and LFDs of the (c*)-test.

![Figure 6](image-url) Asymptotic decrease $I_1$ of the composite test when the observations follow the nominal distributions $F_0 \sim \mathcal{N}(-1, 1)$ and $F_1 \sim \mathcal{N}(1, 2)$, LFDs of the (m)-test, LFDs of the (a)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, LFDs of the (h)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, LFDs of the (c)-test, $\tilde{g}_0$ and $\tilde{g}_1$, with $\epsilon_0 = \epsilon_1 = 0.01$, and LFDs of the (c*)-test.
ratio between LFDs obtained by single sample robust tests. The robustness of the composite model strictly depends on the robustness of each single model: the sequential (m)-test and the sequential (h)-test. If one of them fails to be minimax robust, then the composite model is not minimax robust either. This makes the analysis of the test of robustness for the sequential (m)-test and the sequential (h)-test general enough to have conclusions about the composite test. In the sequel, Monte-Carlo simulations have been performed with 10^5 samples. The threshold space \((\ln t_1, \ln t_a) \in \mathbb{R}^- \times \mathbb{R}^+\) is first cropped to \([-6, 0] \times [0, 6]\) and then discretized with a step parameter of 0.01 in both directions, leading to 60 \times 60 pairs of \((\ln t_1, \ln t_a)\). The nominal distributions are selected to be the mean and variance shifted Gaussian distributions as before. For \(\varepsilon = \varepsilon_0 = \varepsilon_1 = 0.01\), the LFDs of the (m)-test \((\hat{g}_0, \hat{g}_1)\) and the (a)-test \((\hat{g}_0, \tilde{g}_1)\) are determined by solving (15), (17) and (34). Accordingly, the likelihood ratio is formed by \(l = g_1/g_0\) or \(l = \tilde{g}_1/\hat{g}_0\). The tests considered are \(S_n = \sum_{i=1}^{n} l(Y_i)\) and \(\tilde{S}_n = \sum_{i=1}^{n} \tilde{l}(Y_i)\) where every \(Y_i\) is distributed either as \(g_0\) or \(g_0\) under \(\mathcal{H}_0\) and either \(\tilde{g}_1\) or \(\hat{g}_1\) under \(\mathcal{H}_1\). For every pair of thresholds \((\ln t_1, \ln t_a)\), the sequential test is run and the false alarm probability, miss detection probability and expected number of samples under \(\mathcal{H}_0\) and \(\mathcal{H}_1\) are calculated. Figure 10 illustrates the ratio of the false alarm probability \(\alpha^a_m\) to the false alarm probability \(\alpha^m_m\). Clearly, the performance of \(\tilde{S}_n\) for \(Y_i \sim \hat{g}_0\) degrades for almost all simulation points if actually \(Y_i \sim g_0\). Figure 11 illustrates similar results for the miss detection probability when the robust test is \(\tilde{S}_n = \sum_{i=1}^{n} \tilde{l}(Y_i)\). Again, the test doesn’t satisfy the bounded error probability condition. Figures 12 and 15 illustrate the same type of simulations for the expected number of samples where similar observations can be made. In conclusion, one can see that the sequential (m)-test is not robust for the error probability as well as for the expected number of samples, whereas the sequential (a)-test is only asymptotically robust for the expected number of samples. The simulation results are in agreement with the theoretical findings. A short comparison of the (m)-test, the (a)-test and the (h)-test are given in Table 1.

VIII. CONCLUSION

A minimax robust hypothesis testing scheme between two composite hypotheses based on the KL-divergence has been proposed. It has been shown that the proposed model reduces to Levy’s robust test (7) when the nominal likelihood ratio is monotone and the nominal probability density functions are symmetric. For comparison purposes, Dabak’s asymptotically robust test (6) has been introduced and the existence of LFDs for this test has been proven without consideration of the geometrical aspects of hypothesis testing. It has been shown that the proposed minimax robust test, the (m)-test, can be combined with Huber’s clipped likelihood ratio test, the (h)-test, in a composite uncertainty model. Hence, the composite test, the (c)-test, provides minimax robustness both for outliers as well as for modeling errors. The existence of LFDs for the composite uncertainty model has also been proven. It has been demonstrated that the proposed composite model reduces to the individual robust test via a suitable choice of the parameters.

To design a robust test for modeling errors, the uncertainty sets
can be constructed by choosing distances different from the KL-divergence. It has been shown that the choice of a distance plays a crucial role in designing the robust tests. Although the robust version of the likelihood ratio test remains the same for many distances, there are examples where this assertion is not true. Among several distances discussed, the symmetrized $\chi^2$ has been found to be more suitable for the design of a robust hypothesis test if the tail structures of the nominal distributions are needed to be roughly preserved. It has been also shown that the maximum robustness parameters are bounded from above. Both for the (m)-test as well as for the (h)-test, the problem of determining the maximum robustness parameters is proven to be a convex optimization problem, and therefore the related equations can be solved by a polynomial time algorithm.

Next, the single sample robust tests have been extended to fixed sample size tests. Cramér’s theorem has been adopted to characterize the asymptotic behavior of the robust tests. Interestingly, it has been found that the formulation of the asymptotic decrease rate of the error probability for the fixed sample size test coincides with the formulation to determine the maximum robustness parameters for the (m)-test. Later, single sample robust tests have been extended to the sequential
hypothesis test. The minimax properties of the considered robust tests have either been proven or disproven analytically or with simulations. Finally, we have justified that the proposed composite model is applicable for robust estimation problems. Various simulation results show the agreement with theoretical findings.

| TABLE I |
| COMPARISON BETWEEN THE ROBUST TESTS |
| (m)-test | (a)-test | (h)-test |
|---|---|---|
| Unique LFDs | Yes | Yes | No [5] |
| Unique test | Yes | Yes | Yes |
| Limiting test | Soft sign test | Like. ratio test | Sign test |
| Suitable for | Model. errors | Model. errors | Outliers |
| Non-linear equations | Two coupled | Two distinct | Two distinct |
| Number of samples | \( n = 1 \) | \( n \to \infty \) | \( 1 < n < \infty \) |
| Fixed sample size test | Not robust | Asymp. rob. [6] | Robust |
| Sequential test, \( (\alpha, \beta) \) | Not robust | Not robust | Robust |
| Sequential test, \( E[N] \) | Not robust | Asymp. rob. | Asymp. rob. |

ACKNOWLEDGMENT

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

REFERENCES

[1] S. M. Kay, Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory. Prentice Hall PTR, Jan. 1998.
[2] B. C. Levy, Principles of Signal Detection and Parameter Estimation, 1st ed. Springer Publishing Company, Incorporated, 2008.
[3] J. D. Gibson and J. L. Melsa, Introduction to nonparametric detection with applications, ser. Mathematics in science and engineering. New York, San Francisco, London: Academic Press, 1975. [Online]. Available: http://opac.inria.fr/record=b1128538
[4] P. J. Huber, Robust statistics. Wiley New York, 1981.
[5] ——, “A robust version of the probability ratio test,” Ann. Math. Statist., vol. 36, pp. 1753–1758, 1965.
[6] ——, “Robust sequential testing of hypotheses on discrete probability distributions,” Austrian Journal of Statistics, vol. 34, no. 2, pp. 153–162, 2005.
[7] ——, “Robust hypothesis testing with composite distances,” in Proceedings of the Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore, MD, May 1994, pp. 73–77.
[8] ——, “Robust hypothesis testing with squared Hellinger distance,” in Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 2014, p. 1083.
[9] ——, “Minimax tests and the Neyman-Pearson lemma for capacities,” Ann. Statistics, vol. 1, pp. 251–263, 1973.
[10] ——, “Robust detection under communication constraints,” in Proceedings of the IEEE 14th International Workshop on Advances in Wireless Communications (SPAWC), Vancouver, Canada, Jun. 2013, p. 109.
[11] ——, “Robust detection testing with squared Hellinger distance,” in Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 2014, p. 1083.
[12] ——, “Robust hypothesis testing for modeling errors,” in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 2013, pp. 5514–5518.
[13] ——, “Robust hypothesis testing with squared Hellinger distance,” in Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 2014, p. 1083.
[14] ——, “Minimax tests and the Neyman-Pearson lemma for capacities,” Ann. Statistics, vol. 1, pp. 251–263, 1973.
[15] W. Feller, An Introduction to Probability Theory and Its Applications. Wiley, January 1968, vol. 1.