An Approach to Joint Sequential Detection and Estimation with Distributional Uncertainties

Dominik Reinhard, Michael Fauß and Abdelhak M. Zoubir
Signal Processing Group, Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt
Email: {reinhard, fauss, zoubir}@spg.tu-darmstadt.de

***Abstract*—Joint detection and estimation is an important yet little-studied problem that arises in many signal processing applications. In this paper, a sequential and robust solution approach is presented. To design the test fulfilling constraints on the error probabilities and the quality of the estimate, the problem is converted into an unconstrained form and subsequently solved using Linear Programming. To handle model uncertainties, a band model for both hypotheses is used and a concept for determining the pair of least favorable distributions is adopted to devise a robust detection scheme. For the robust estimation, an upper bound of the estimation cost, based on maximizing a Kullback-Leibler divergence, is derived. The resulting test meets the specifications on the error probabilities and the quality of the estimate for every feasible pair of distributions. Numerical results are provided for the pair of least favorable distributions and for a pair of randomly selected distributions.**

***Index Terms*—band model, distributional uncertainties, joint detection and estimation, linear programming, optimization, robustness, sequential analysis**

I. Introduction

In many applications, the problem arises to decide between two hypotheses and, depending on the decision, to estimate some parameters of the underlying distribution. This problem was initially treated by Middleton and Esposito [1] in the late 1960s and revisited by Moustakides in 2011 [2]. Joint detection and estimation is widely used in different areas of signal processing, including speech [3], communication [4] and radar [5].

Sequential analysis is a field of research introduced by Wald in the late 1940s [6]. The idea behind sequential detection is to observe a sequence of samples in order to decide as quickly as possible in favor of one of two hypotheses while fulfilling constraints on the error probabilities. Until the present day, sequential methods have been a topic of continuous research in many fields. Especially for time critical or low power applications, sequential approaches often outperform conventional approaches.

Combining the idea of sequential analysis with the problem of joint detection and estimation leads to a powerful framework that enables one to quickly decide for one hypothesis and, if necessary, accurately estimate unknown parameters of the corresponding distribution. Joint detection and estimation problems of this kind arise in many applications, for example, in digital communication where a communication line is observed and there is the need to establish whether or not a signal is present and, in the latter case, to estimate its mean and/or variance.

Moreover, robust signal processing, i.e., the processing of signals under deviations from the assumed model, is very important for many applications. Often it is the case that the assumptions, for example, Gaussian distributed noise, do not hold in practice or that the distribution of the signal is only known approximately. Therefore, methods have to be developed that can deal with these deviations.

This paper combines these three areas and provides a first approach to sequential joint detection and estimation that handles distributional uncertainties. For this approach, it is assumed that the possible distributions lie inside a known band and the data is identically and independently distributed. A sequential joint detection and estimation approach is developed based on the least favorable instead of the nominal distributions. For the design of the test, a Linear Programming (LP) approach, similar to the one presented in [7], is used.

The paper is structured as follows: In Section II, a detailed description of the problem is given. The design of least favorable distributions, i.e., the distributions which lead to the worst results, is detailed in Section III. The formulation of the problem as a linear program is given in Section IV. Finally, the results are illustrated with numerical examples in Section V.

Regarding the notation: $\mathbf{E}_P$ and $\mathbf{Var}_P$ denote the expected value and the variance of a random variable with respect to the measure P. For all sequences, superscript n denotes the time instant.

II. Problem Formulation

Let $\left(X^1, \dots, X^n\right)$ be a sequence of independent and identically distributed random variables with common distribution P, defined on the probability space $(\Omega, \mathcal{F}, P)$. It is assumed that the distribution P admits a density p with respect to some reference measure μ. The goal in sequential joint detection and estimation is to sequentially perform a binary hypothesis test and to estimate a parameter of the distribution P_1 if one decides in favor of $\mathcal{H}_1$. The simple hypotheses are given by

$$\begin{aligned}\mathcal{H}_0 &: \quad P = P_0,\\ \mathcal{H}_1 &: \quad P = P_1.\end{aligned}$$

When designing a joint detection and estimation scheme, the goal is to minimize the expected run-length of the test while satisfying constraints on the error probabilities as well as on the quality of the estimate. The upper bounds on the error probabilities are denoted by γ_0 and γ_1, whereas the bound on the quality of the estimate is denoted by γ_2. The minimization of the average run-length is performed under a third distribution P. This problem is also known as the modified Kiefer–Weiss problem.

Since in many practical applications the distribution of the data is not known exactly or only a heuristic description of the distribution is available, the presented approach also takes into account deviations from the nominal distributions. Therefore, the so-called band model, introduced by Kassam [8], is used. This model allows the true, unknown probability density to lie inside a band, i.e., $p_i' \leq p_i \leq p_i''$, $i \in \{0,1\}$. Due to this uncertainty, the problem of deciding between two simple hypotheses becomes a problem of deciding between two composite ones. The use of this model also clarifies the need for estimating a parameter, e.g. the mean, of the distribution under the alternative hypothesis.

Mathematically, the design of the test can be formulated as the optimization problem

$$\begin{aligned} &\min_{\Psi,\Phi} \ \mathbf{E}_P\left[\tau\right] && (1)\\ \text{subject to} \quad &\max_{P_0\in\mathcal{P}_0} \ P_0(\Phi^\tau = 1) \leq \gamma_0, \\ &\max_{P_1\in\mathcal{P}_1} \ P_1(\Phi^\tau = 0) \leq \gamma_1, \\ &\max_{P_1\in\mathcal{P}_1} \ \mathbf{Var}_{P_1}\left[\hat{\theta}\right] \leq \gamma_2, \end{aligned}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ denote the uncertainty sets of feasible distributions, and Φ and Ψ denote the decision and stopping rules of the test, respectively. More precisely, $\Phi^n(x^1,\ldots,x^n) = i$ denotes a decision for $\mathcal{H}_i$, $i \in \{0,1\}$, at time instant n and $\Psi^n(x^1,\ldots,x^n) = 1$ or $\Psi^n(x^1,\ldots,x^n) = 0$ denote the decisions to stop or continue the test at time instant n. Both functions are dependent on the observations $x^1,\ldots,x^n$. The stopping time τ of the test is defined as $\tau = \min\{n \geq 0 : \Psi^n = 1\}$. The estimator for the unknown parameter is denoted by $\hat{\theta}$.

III. Selection of the Least Favorable Distributions

The use of the word robust varies widely in the literature. Hence, we first clarify our definition of robustness before going into details about how to choose the least favorable distributions.

As mentioned before, we allow the distributions under the null hypothesis and the alternative to lie inside a band. In this work, we want to construct a test that fulfills the constrains on the error probabilities and the quality of the estimate for every pair of distributions lying inside the specified bands.

In order to satisfy the constraints on the error probabilities, a pair of distributions has to be found that are least separable or *least favorable*. In other words, the distributions under the null hypothesis and the alternative should be as similar as possible. For the fixed sample size (FSS) test, it has been shown that a pair of distributions is least favorable if it jointly minimizes all f-divergences. See, for example, [8]–[10].

In the FSS scenario, the optimal minimax test is a likelihood ratio test that uses the least favorable distributions instead of the nominal ones [11]. The design of a minimax optimal procedure for joint sequential detection and estimation is more challenging and has not been studied yet. Although replacing the nominal distributions by the least favorable ones does not yield a strictly optimal procedure, this method is used here as a first step towards a minimax scheme.

In this work, the implicit characterization of the pair of least favorable distributions stated in [10, Eq. 13] is used. It is given by

$$\begin{aligned} q_0 &= \min\left\{p_0'', \max\left\{c_0\left(\alpha q_0 + q_1\right), p_0'\right\}\right\} \\ q_1 &= \min\left\{p_1'', \max\left\{c_1\left(\alpha q_1 + q_0\right), p_1'\right\}\right\} \ , \end{aligned} \tag{2}$$

where q_0 and q_1 denote the least favorable distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, $c_{\{0,1\}} \in \left[0, \frac{1}{\alpha}\right]$ and α is a positive constant. For the numerical calculation of q_0 and q_1, we use the iterative construction algorithm [10, Tab. I].

Having robustified the detection of the joint problem, the estimation subproblem is to be solved. When referring to robust estimation, one usually means an estimator which is robust itself, for example, the well-known M-Estimator [12]. In this work, instead of making the estimator itself robust, a distribution is sought that provides an upper bound on the variance of the estimator $\mathbf{Var}\big[\hat{\theta}\big]$. This is done in two independent steps, which are illustrated below, using the example of the sample mean. An extension to other estimators is possible, but it is beyond the scope of this paper.

Using the sample mean as an estimator, it holds that

$$\mathbf{Var}_{P_1}\left[\hat{\theta}\right] = \mathbf{Var}_{P_1}\left[\frac{1}{\tau}\sum_{n=1}^{\tau} x^n\right] = \mathbf{E}_{P_1}\left[\frac{\sigma^2}{\tau}\right], \tag{3}$$

where σ^2 denotes the variance of P_1. A distribution is least favorable with respect to the estimator variance if it maximizes the expected value on the right-hand side of Eq. (3). However, determining this maximum is a formidable task since σ^2 and τ are coupled and the latter additionally depends on the stopping rule. In order to make the problem more tractable and still guarantee that the accuracy constraints on the estimate are met, we resort to an upper bound on Eq. (3). This bound is derived by independently maximizing the numerator and minimizing the denominator in Eq. (3), i.e., by determining *two* least favorable distributions: one with maximum variance σ^2 and one which minimizes the expected run-length under the alternative.

The maximum variance among all distributions that lie within the density band corresponding to $\mathcal{H}_1$ is in the following denoted by κ^2. To calculate the distribution that achieves this maximum, the following maximization problem has to be solved

$$\kappa^2 = \max_f \left(\int x^2 f d\mu(x) - \left(\int x f d\mu(x)\right)^2\right) \tag{4}$$

$$\text{s.t.} \quad f \leq p_i'' \ , \quad f \geq p_i' \ , \quad \int f d\mu = 1,$$

which is a convex problem and can thus be solved numerically with off-the-shelve solvers.

In a second step, a distribution has to be found that minimizes the expected run-length of the test. Intuitively speaking, this distribution needs to be chosen such that it is least similar to the distribution under the null hypothesis. According to Wald [6, Eq. 3:68], the expected run-length under the alternative hypothesis of a sequential test is approximately proportional to the inverse of the Kullback–Leibler divergence of P_1 and P_0, i.e.,

$$E_{P_1}[\tau] \propto \frac{1}{\mathbf{E}_{P_1}\left[\log \frac{p_1}{p_0}\right]} =: \frac{1}{D_{\mathrm{KL}}(p_1 \| p_0)} . \tag{5}$$

From this approximation, it can be seen that minimizing the expected run-length is approximately equivalent to maximizing the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(p_1\|p_0)$. This results in the following maximization problem

$$\max_{q} \; D_{\mathrm{KL}}(q\|p_0) \tag{6}$$
$$\text{s.t.} \quad q \leq p_1'', \quad q \geq p_1', \quad \int q \, d\mu = 1,$$

where

$$D_{\mathrm{KL}}(q\|p_0) = \int q(x) \log\left(\frac{q(x)}{p_0(x)}\right) d\mu(x). \tag{7}$$

Since Problem (6) is nonconvex, it cannot be solved using standard convex optimization techniques. Here, we suggest the use of a projected gradient algorithm.

In order to apply a projected gradient algorithm to the problem at hand, the densities are first discretized. The gradient with respect to q is then calculated as

$$\nabla_q D_{\mathrm{KL}}(q\|p_0) = 1 + \log \frac{q}{p_0}. \tag{8}$$

The projection step is performed by solving

$$\mathrm{Pr}_{\mathcal{P}_1}(q) = \min\left\{p_1'', \max\left\{cq, p_1'\right\}\right\}, \tag{9}$$

for the constant c such that the result is a valid density. This procedure is repeated until the solution converges.

Given the two least favorable distributions, the constraint for the quality of the estimate is replaced by

$$E_Q\left[\frac{\kappa^2}{\tau}\right] \leq \gamma_2 , \tag{10}$$

where $E_Q[\cdot]$ is the expected value with respect to the probability distribution Q that solves (6) and κ^2 is defined in Eq. (4).

By using the pair of least favorable distributions obtained from Eq. (2) for the constraints on the error probabilities, and the upper bound for the estimation constraint stated in Eq. (10), the final optimization problem becomes

$$\begin{aligned} &\min_{\Psi,\Phi} \; \mathbf{E}_Q[\tau] && (11) \\ \text{subject to} \quad & Q_0(\Phi^\tau = 1) \leq \gamma_0, \\ & Q_1(\Phi^\tau = 0) \leq \gamma_1, \\ & E_Q\left[\frac{\kappa^2}{\tau}\right] \leq \gamma_2, \end{aligned}$$

where the expected run-length of the test is now minimized under distribution Q.

IV. Sequential Joint Detection and Estimation as a Linear Program

To obtain an optimal test for the joint detection and estimation task, the constrained minimization Problem (11) is converted to an unconstrained one using the method of Lagrange multipliers. This resulting problem can then be treated as an optimal stopping problem for which Novikov provided optimal stopping and decision rules [13]. Using these rules and following the line of arguments in [7], [13] yields an optimal test, the cost function of which is defined recursively by

$$\rho_\lambda^n(z) = \min\left\{g_\lambda^n(z), 1 + \int \rho_\lambda^{n+1} dH_z\right\}, \tag{12}$$

where

$$g_\lambda^n(z) := \min\left\{\lambda_0 z_0, \lambda_1 z_1\right\} + \lambda_2 \frac{\kappa^2}{n} . \tag{13}$$

Here the superscript denotes the time instant, $z = (z_0, z_1)$ is a tuple of likelihood ratios, $\lambda = (\lambda_0, \lambda_1, \lambda_2)$ a triplet of Lagrange multipliers and the probability measure H_z is defined as

$$H_z(B) = P\left(\left\{\left(z_0 \frac{q_0(x)}{q(x)}, z_1 \frac{q_1(x)}{q(x)}\right) \in B\right\}\right), \tag{14}$$

where B is an element of the Borel σ-algebra on $[0,\infty)^2$. In Eq. (12), the term g_λ^n corresponds to the cost for stopping at time instant n, and the term $1 + \int \rho_\lambda^{n+1} dH_z$ to the cost for continuing. For the sake of a more compact notation, the latter is denoted by d_λ^n in the sequel. The test stops if $g_\lambda^n < d_\lambda^n$ and decides in favor of $\mathcal{H}_0$ if $\lambda_0 z_0 \leq \lambda_1 z_1$, otherwise in favor of $\mathcal{H}_1$. Opposed to [7], a truncated sequential procedure is considered in this work, which means that the test is forced to stop at time instant N so that $\rho_\lambda^N = g_\lambda^N$. Minimizing the expected run-length of the test is then equivalent to minimizing $\rho^0(1,1)$ [7], where $z = (1,1)$ is the initial state of the test statistic.

The question how to choose the Lagrange multiplier such that certain constraints on the error probabilities are fulfilled is addressed in [7], where an approach based on linear programming is suggested. Since the sequential joint detection and estimation problem has a similar form, this method can be applied to the joint problem as well. A rigorous proof of this extension is omitted for brevity, but can be found in [14].

In order to obtain the Lagrange multipliers, the Lagrangian dual of (11) is maximized, which results in the following maximization problem

$$\begin{aligned} &\max_{\lambda > 0} \; \rho_\lambda^0(1,1) - \lambda_0\gamma_0 - \lambda_1\gamma_1 - \lambda_2\gamma_2 , && (15) \\ \text{subject to} \quad & \rho_\lambda^n(z) = \min\left\{g_\lambda^n(z), 1 + \int \rho_\lambda^{n+1} dH_z\right\} , \\ & \rho_\lambda^N(z) = g_\lambda^N(z) . \end{aligned}$$

Using arguments similar to those in [7], it can be shown that a test with Lagrange multipliers chosen according to (15) meets the constraints on the error probabilities and the estimation quality exactly. Hence, the bounds on the error probabilities and on the estimation quality are also called target error probabilities and target estimation cost, respectively.

Adding the sequence of functions ρ_λ^n to the set of free variables yields the maximization problem that is used for the design of the test:

$$\max_{\lambda>0,\rho^n\in\mathcal{L}} \rho_\lambda^0(1,1) - \lambda_0\gamma_0 - \lambda_1\gamma_1 - \lambda_2\gamma_2, \tag{16}$$

$$\text{s.t.} \quad \rho_\lambda^n(z) \le \min\left\{\lambda_0 z_0 + \lambda_2\frac{\kappa^2}{n}, \lambda_1 z_1 + \lambda_2\frac{\kappa^2}{n}, 1 + \int \rho_\lambda^{n+1} dH_z\right\},$$

$$\rho_\lambda^N(z) \le \min\left\{\lambda_0 z_0 + \lambda_2\frac{\kappa^2}{N}, \lambda_1 z_1 + \lambda_2\frac{\kappa^2}{N}\right\},$$

where $\mathcal{L}$ denotes the set of all H_z integrable functions. Since the objective of the maximization problem is concave in λ and ρ^n, this problem can be solved using convex optimization techniques. It can be shown that the optimal value is the expected run-length of the test. The proof is given in [14].

V. Numerical Results

In this section, a numerical example is given to illustrate the proposed approach for robust sequential joint detection and estimation. Monte Carlo results are provided for the case when the observations are generated under the least favorable distributions, as defined in Section III, and under a pair of distributions that was chosen randomly from the respective density bands. Before presenting the numerical results, the necessary quantities, which were used for the test design and the Monte Carlo simulation are introduced.

Since the values of the likelihood ratio usually cover a wide numerical range, which often results in numerical inaccuracies, the log-likelihood ratio is used instead. To numerically solve Problem (16), the log-likelihood ratios were sampled uniformly on $[-6,6]\times[-6,6]$ with a step size of 0.2. To enforce numerical stability, the following term was added to the optimization problem

$$\varepsilon\frac{1}{N+1}\sum_{n=0}^{N}\int \rho_\lambda^n d\mu, \tag{17}$$

where ε is a small positive constant, see [7, App. F] for details. In the simulation presented in this work, this constant is set to $\varepsilon = 10^{-4}$. Furthermore, the maximum number of samples allowed for the truncated test is $N = 30$. For the Monte Carlos simulations, 20,000 samples are generated under each distribution.

Truncated Gaussian distributions were used as nominal distributions for the test design. The means of the Gaussian distributions were chosen as $\mu_1 = -\mu_0 = 0.5$ and their variance was set to be $\sigma^2 = 1$. The distributions were sampled on the interval $[-2.5, 2.5]$ with a step size of 10^{-2}. The band $0.9p_i \le q_i \le 1.1p_i$, $i\in\{0,1\}$, was used to model distributional uncertainties. The band model and the resulting set of least favorable distributions are depicted in Fig. 1 and Fig. 2, respectively.

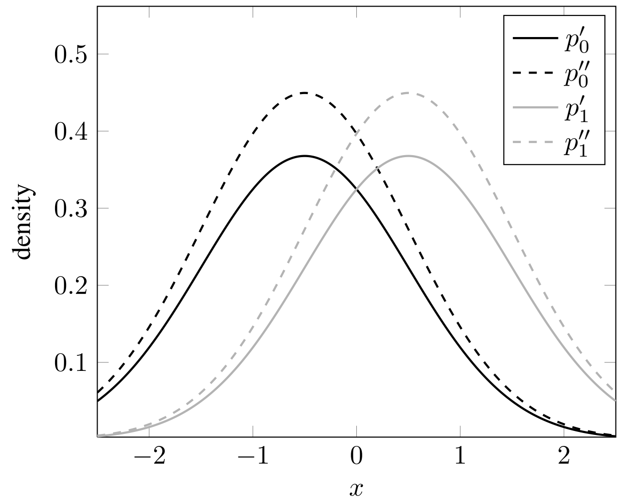

Figure 1. Band model

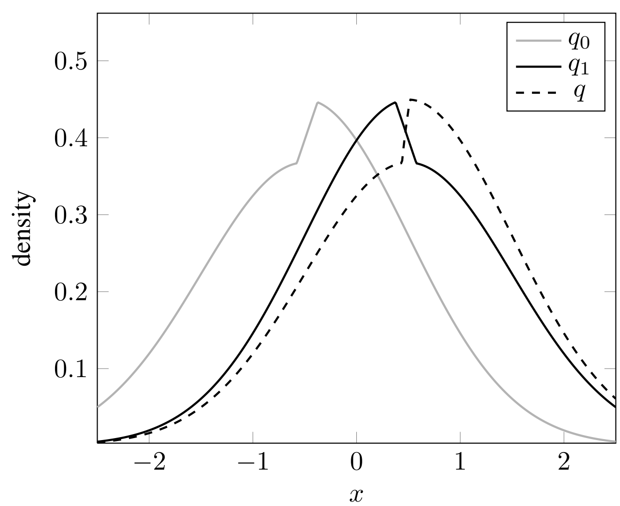

Figure 2. Densities of least favorable distributions

For the design of the test, first, the least favorable distributions for detection, q_0 and q_1, were determined according to Eq. (2). Subsequently, the least favorable distributions for estimation were determined according to the maximizations in Eq. (4) and Eq. (6). In order to solve (6) for q, it was further assumed that $p_0 = q_0$, i.e., the accuracy is bounded under the least favorable distribution. The expected run-length is accordingly minimized under $p = q$.

It can be seen that the distributions q_0 and q_1 are the ones which are most difficult to separate since they are very close together, whereas the distribution q can be separated from q_0 much more easily.

By inspection of the numerical results listed in Table I, one can see that all constraints are fulfilled when running a Monte

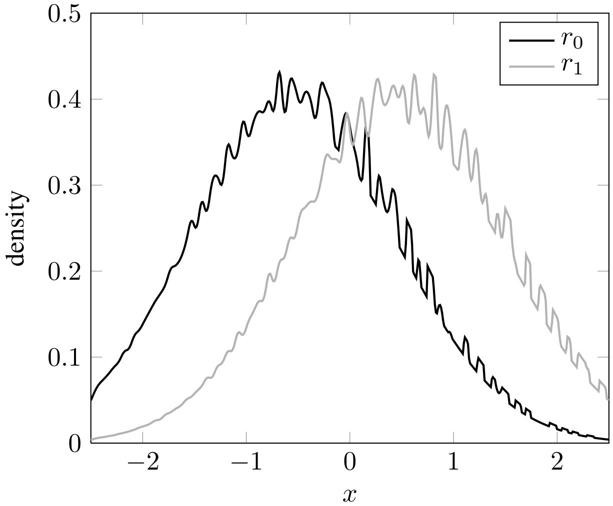

Figure 3. Densities of randomly chosen distributions

Table I
NUMERICAL RESULTS

	least favorable distributions	random distributions	target
type-I error	0.045	0.022	0.05
type-II error	0.046	0.032	0.05
$E\left[\frac{\kappa^2}{\tau}\right]$	0.093	0.096	0.10
$E[\tau]$	14.34	14.13	9.68
λ_0	6.81	6.81	–
λ_1	2.82	2.82	–
λ_2	88.08	88.08	–

Carlo simulation with the set of least favorable distributions. It is noteworthy that the expected run-length of the test, which is provided by the LP approach, differs 4 to 5 samples from the results of the Monte Carlo simulations. This can be explained by the fact the test is designed under the distribution q, which means that the expected run-length is minimized under the distribution q and, therefore, the expected run-length under all other distributions is larger. By having a look at Fig. 2, one can see that the distribution q differs significantly from the least favorable distributions under both hypothesis. This is the case since q is chosen such that the expected run-length is minimized.

The designed test was applied to a second data set, where the data was generated under randomly selected distributions inside the band model. These distributions are shown in Fig. 3. By inspection of the third column of Table I, one can see that all constrains are fulfilled. The difference in the expected run-length of the test can be explained in the same way as for the set of least favorable distributions.

The results presented in this section show that the proposed approach can provide tests that achieve a predefined performance even under distributional uncertainties.

VI. CONCLUSION AND OUTLOOK

This paper provides a first attempt at robust joint sequential detection and estimation. The resulting test fulfills constraints on the error probabilities and on the quality of the estimate under distributional uncertainties. By adopting methods from the fixed sample size case, an upper bound for the error probabilities was found such that the constraints are fulfilled for all feasible distributions. To ensure the performance of the estimator, two least favorable distributions were used to formulate an upper bound on the estimation quality. These least favorable distributions maximize the variance and minimize the expected run-length, respectively.

The test is not optimal in the minimax sense, but it is guaranteed to achieve a pre-specified performance for all feasible distributions. Deriving an optimal minimax scheme is a challenge for further research. Moreover, the very strict assumption of identically and independently distributed random variables should be relaxed in the next steps. Finally, the class of possible estimators has to be extended such that all necessary parameters of the underlying distribution can be estimated.

REFERENCES

[1] D. Middleton and R. Esposito, “Simultaneous optimum detection and estimation of signals in noise,” *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 434–444, May 1968.

[2] G. Moustakides, “Optimum joint detection and estimation,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, July 2011, pp. 2984–2988.

[3] H. Momeni, H. Abutalebi, and A. Tadaion, “Joint detection and estimation of speech spectral amplitude using noncontinuous gain functions,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1249–1258, Aug 2015.

[4] C. Cozzo and B. Hughes, “Joint detection and estimation in space-time coding and modulation,” in *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, vol. 1, Oct 1999, pp. 613–617 vol.1.

[5] A. Tajer, G. Jajamovich, X. Wang, and G. Moustakides, “Optimal joint target detection and parameter estimation by MIMO radar,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 1, pp. 127–145, Feb 2010.

[6] A. Wald, *Sequential Analysis*, 2nd ed. New York: Wiley, 1948.

[7] M. Fauß and A. M. Zoubir, “A linear programming approach to sequential hypothesis testing,” *Sequential Analysis*, vol. 34, no. 2, pp. 235–263, 2015.

[8] S. A. Kassam, “Robust hypothesis testing for bounded classes of probability densities,” *IEEE Transactions on Information Theory*, vol. 27, pp. 242–247, 1981.

[9] P. J. Huber, “A robust version of the probability ratio test,” *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[10] M. Fauß and A. M. Zoubir, “Old bands, new tracks—Revisiting the band model for robust hypothesis testing,” 2015, under review for publication in the *IEEE Transactions on Signal Processing*. [Online]. Available: http://arxiv.org/abs/1510.04524

[11] H. Komiya, “Elementary proof for Sion’s minimax theorem,” *Kodai mathematical journal*, vol. 11, no. 1, pp. 5–7, 1988.

[12] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[13] A. Novikov, “Optimal sequential tests for two simple hypotheses,” *Sequential analysis*, vol. 28, no. 2, pp. 188–217, 2009.

[14] D. Reinhard, “Sequential joint detection and estimation,” Master’s thesis, Technische Universität Darmstadt, Germany, 2016.