

”This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.”

A BAYESIAN APPROACH TO SPECTRUM SENSING, DENOISING AND ANOMALY DETECTION

Erik Axell and Erik G. Larsson

Department of Electrical Engineering (ISY), Linköping University, 581 83 Linköping, Sweden
 {axell, erik.larsson}@isy.liu.se

ABSTRACT

This paper deals with the problem of discriminating samples that contain only noise from samples that contain a signal embedded in noise. The focus is on the case when the variance of the noise is unknown. We derive the optimal soft decision detector using a Bayesian approach. The complexity of this optimal detector grows exponentially with the number of observations and as a remedy, we propose a number of approximations to it. The problem under study is a fundamental one and it has applications in signal denoising, anomaly detection, and spectrum sensing for cognitive radio. We illustrate the results in the context of the latter.

Index Terms— spectrum sensing, denoising, anomaly detection

1. INTRODUCTION

This paper deals with the problem of discriminating samples that contain only noise, from samples that contain a signal embedded in noise. More precisely, out of a total of M observations y_i , $i = 1, \dots, M$, we want to determine which samples that are realizations of a noise process and which samples that contain a signal corrupted by additive noise. If the distribution of the noise is known and the observations y_i are independent, then an energy detector is essentially optimal, and it consists of comparing each $|y_i|$ to a threshold. The focus of our work is on the case when the noise variance is *unknown* (but the same for all observations). In this case, the observations y_i become correlated and the optimal detector *cannot* be implemented by simple thresholding of $|y_i|$. We derive the optimal detector in a Bayesian framework, and devise a computationally efficient approximation of it.

The main motivating application for the problem under study is spectrum sensing for cognitive radio. The key problem in cognitive radio is to find “spectrum holes”, and to do this one must detect very weak signals. Typically, multiple bands are scanned simultaneously [1, 2], and y_i is then the observation in the i th band. In spectrum sensing applications, one may also wish to combine many independent spectrum measurements at a fusion center [3, 4]. To facilitate this, the detectors should deliver reliability information on their decisions (“soft decisions”). What is important is then not only to take individual, hard decisions on whether a signal is present in a specific band i , but to determine the a posteriori probability that there is a signal present in band i , given all available observations.

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216076. This work was also supported in part by the Swedish Research Council (VR) and the Swedish Foundation for Strategic Research (SSF). E. Larsson is a Royal Swedish Academy of Sciences (KVA) Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation.

Two other important applications of the problem we study here are denoising of data (e.g., see [5]) and detection of anomalies in time-series [6]. The problem also has connections to sparse signal modeling. In particular, it can be viewed as a special case of linear regression with a sparse coefficient vector [7] and with an identity matrix as regression matrix. The main contribution of this paper relative to [7] is that we deal systematically (in a Bayesian framework) with the case of unknown noise variance, and that we derive a soft output detector. We also provide illustrations in the context of cooperative spectrum sensing for cognitive radio.

2. PROBLEM FORMULATION

We assume that we have M independent observations y_i , $i = 1, 2, \dots, M$. Each observation contains noise n_i only, with probability p , and a signal x_i embedded in noise with probability $1 - p$. That is:

$$\begin{cases} y_i = n_i, & \text{with probability } p, \\ y_i = x_i + n_i, & \text{with probability } 1 - p. \end{cases}$$

We assume that the noise and signal are independent zero-mean Gaussian random variables with different variances, more precisely: $n_i \sim N(0, \sigma^2)$, and $x_i \sim N(0, \rho^2)$. The noise variance σ^2 and the signal variance ρ^2 are assumed to be unknown. (If they were known, the optimal detector would simply consist of M independent binary hypothesis tests; see also the end of Section 3.)

We define the following 2^M hypotheses:

$$\begin{cases} H_0 : & y_1 = n_1, y_2 = n_2, \dots, y_M = n_M, \\ H_1 : & y_1 = x_1 + n_1, y_2 = n_2, \dots, y_M = n_M, \\ H_2 : & y_1 = n_1, y_2 = x_2 + n_2, y_3 = n_3, \dots, y_M = n_M, \\ \vdots & \\ H_{2^M-2} : & y_1 = x_1 + n_1, y_2 = x_2 + n_2, \dots, \\ & y_{M-1} = x_{M-1} + n_{M-1}, y_M = n_M, \\ H_{2^M-1} : & y_1 = x_1 + n_1, y_2 = x_2 + n_2, \dots, \\ & y_M = x_M + n_M. \end{cases}$$

We assume that the signal presence is independent between all observations. Thus, we obtain the following a priori probabilities:

$$\begin{cases} P(H_0) = p^M, \\ P(H_1) = P(H_2) = \dots = P(H_M) = (1-p)p^{M-1}, \\ \vdots \\ P(H_{2^M-M-1}) = \dots = P(H_{2^M-2}) = (1-p)^{M-1}p, \\ P(H_{2^M-1}) = (1-p)^M. \end{cases}$$

For each hypothesis, H_i , let S_i be the set of observation indices for which signal is present:

$$\begin{cases} S_0 = \emptyset, S_1 = \{1\}, S_2 = \{2\}, S_M = \{M\}, \dots, \\ S_{M+1} = \{1, 2\}, S_{M+2} = \{1, 3\}, \dots, \\ S_{2M-2} = \{1, 2, \dots, M-1\}, S_{2M-1} = \{1, 2, \dots, M\}. \end{cases}$$

Then the likelihood of the received sequence $y = (y_1, y_2, \dots, y_M)$ under hypothesis H_i , and for given σ and ρ , is

$$P(y|H_i, \sigma, \rho) = \prod_{k \in \bar{S}_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} |y_k|^2\right) \times \prod_{k \in S_i} \frac{1}{\sqrt{2\pi}(\sigma^2 + \rho^2)} \exp\left(-\frac{1}{2(\sigma^2 + \rho^2)} |y_k|^2\right).$$

3. OPTIMAL DETECTOR

Using Bayes rule, we can write the a posteriori probability of hypothesis H_i given y , σ and ρ , as

$$P(H_i|y, \sigma, \rho) = \frac{P(y, H_i|\sigma, \rho)}{P(y|\sigma, \rho)} = \frac{P(y|H_i, \sigma, \rho)P(H_i|\sigma, \rho)}{P(y|\sigma, \rho)}.$$

The hypotheses H_i are assumed to be independent of the variances σ and ρ . Hence, $P(H_i|\sigma, \rho) = P(H_i)$.

Ultimately we are typically interested in the probability of the event that a signal is present in the i th observation, given y . Let Ω_i denote this event. The probability of Ω_i , given the observation y , can be written

$$P(\Omega_i|y) = \sum_{k:i \in S_k} P(H_k|y) = \frac{\sum_{k:i \in S_k} P(y|H_k)P(H_k)}{\sum_{m=0}^{2^M-1} P(y|H_m)P(H_m)}, \quad (1)$$

where $P(y|H_k)$ is $P(y|H_k, \sigma, \rho)$ with σ and ρ eliminated (via marginalization, or using approximations such as inserting estimates of σ and ρ). In the following sections we discuss how to deal with this marginalization problem.

Often one is interested in combining decisions on Ω_i made by different sensors. An important example is cooperative spectrum sensing (see discussion in Section 1). To facilitate such combining we define the soft decision value for the i th observation (i th band) as the log-likelihood ratio

$$\lambda_i \triangleq \log \left(\frac{P(\Omega_i|y)}{P(\bar{\Omega}_i|y)} \right) = \log \left(\frac{\sum_{k:i \in S_k} P(y|H_k)P(H_k)}{\sum_{k:i \in \bar{S}_k} P(y|H_k)P(H_k)} \right). \quad (2)$$

If there are C , say, independent cooperating sensors then we can obtain a soft decision value $\lambda_{c,i}$ for each band i from each cooperating sensor c . If each sensor observes the same true hypothesis H_k but the noise and signal random variables are independent across the sensors, then it is optimal to add the log-likelihood ratios in (2) at the fusion center. (This also assumes that the soft decision values are transmitted error-free to the fusion center.) Hard decisions on whether a signal is present in the i th observation or not, are then taken at the fusion center based on

$$\Lambda_i \triangleq \sum_{c=1}^C \lambda_{c,i} \underset{\text{no signal in band } i}{\overset{\text{signal in band } i}{\geq}} \mu, \quad (3)$$

where μ is a detection threshold.

As a benchmark for comparison, we give the optimal detector

when ρ and σ are known. The observations y_i will then be mutually independent and the 2^M composite hypothesis test decouples to M independent binary hypothesis tests, one for each i .

Equation (2) becomes

$$\lambda_i = \log \left(\frac{\frac{1}{\sqrt{2\pi}(\sigma^2 + \rho^2)} \exp\left(-\frac{1}{2(\sigma^2 + \rho^2)} |y_i|^2\right) \cdot (1-p)}{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} |y_i|^2\right) \cdot p} \right) = |y_i|^2 \frac{\rho^2}{2\sigma^2(\sigma^2 + \rho^2)} + \frac{1}{2} \log \left(\frac{\sigma^2}{\sigma^2 + \rho^2} \right) + \log \left(\frac{1-p}{p} \right), \quad (4)$$

which is then used in (3) to take decisions.

4. DETECTOR FOR UNKNOWN ρ, σ

Next we consider the case where the ρ and σ are unknown. We propose two ways of dealing with the fact that these variances are unknown: estimation and marginalization.

4.1. Estimation of ρ, σ using prior knowledge

Suppose that we know, a priori, that m of the M observations contain only noise. Then, we can use this information to estimate the noise variance σ^2 from the m smallest observations:

$$\widehat{\sigma^2} = \frac{1}{m} \sum_{m \text{ smallest}} |y_k|^2. \quad (5)$$

Furthermore, suppose that we know that s of the M observations contain signal plus noise. In a similar manner, we could then estimate the signal-plus-noise variance $\sigma^2 + \rho^2$ from the s largest observations:

$$\widehat{\sigma^2 + \rho^2} = \frac{1}{s} \sum_{s \text{ largest}} |y_k|^2. \quad (6)$$

If we know the a priori probability of finding a signal in an observation, $1-p$, then the number of observations that contain only noise is binomially distributed with mean pM . A natural choice is then to use the $m = pM$ smallest observations to compute $\widehat{\sigma^2}$. Similarly, we can use the $s = (1-p)M$ largest observations to compute $\widehat{\sigma^2 + \rho^2}$.

When both ρ and σ are estimated and we treat them as given, by inserting $\widehat{\sigma^2}$, $\widehat{\sigma^2 + \rho^2}$ into (2)), the problem decouples just as when the variances are known. Hence, the optimal test for estimated variances consists of using (4) with $\widehat{\sigma^2}$, $\widehat{\sigma^2 + \rho^2}$ inserted in lieu of σ^2 , $\sigma^2 + \rho^2$.

4.2. Elimination of σ via marginalization

The estimation approach of Section 4.1 may be undesirable for several reasons. For example, one may not accurately know p . An alternative is then to postulate a prior for σ and eliminate σ from (2) by marginalization. We will use a Gamma distribution as prior for $\gamma \triangleq 1/\sigma^2$. More precisely, we take $\gamma \sim \text{Gamma}(c, \theta)$, so that

$$P(\gamma) = \gamma^{c-1} \frac{\exp(-\gamma/\theta)}{\theta^c \Gamma(c)}.$$

The motivation for assuming the Gamma distribution is that when $c\theta = 1$ and $c \rightarrow 0$, it becomes non-informative and scaling invariant

[8]. This means that in the limit of $c \rightarrow 0$, $\log(\gamma)$ has a flat distribution. Another benefit is that the marginalization with respect to σ can be computed in closed form. To proceed, assume that σ^2 and $\sigma^2 + \rho^2$ are independent, and let $\beta \triangleq 1/(\sigma^2 + \rho^2)$. Then

$$P(y|H_i, \rho) = \int_0^\infty P(y|H_i, \gamma, \rho)P(\gamma)d\gamma = \int_0^\infty \prod_{k \in \bar{S}_i} \sqrt{\frac{\gamma}{2\pi}} \exp(-\frac{1}{2}\gamma|y_k|^2) \cdot \prod_{l \in S_i} \sqrt{\frac{\beta}{2\pi}} \exp(-\frac{1}{2}\beta|y_l|^2) \times \gamma^{c-1} \frac{\exp(-\gamma/\theta)}{\theta^c \Gamma(c)} d\gamma = \frac{\Gamma(c + |\bar{S}_i|/2)}{(2\pi)^{|\bar{S}_i|/2} \theta^c \Gamma(c) \left(\frac{1}{2} \sum_{k \in \bar{S}_i} |y_k|^2 + \frac{1}{\theta}\right)^{c+|\bar{S}_i|/2}} \times \prod_{l \in S_i} \sqrt{\frac{\beta}{2\pi}} \exp(-\frac{1}{2}\beta|y_l|^2),$$

where $|\bar{S}_i|$ denotes the number of elements of the set \bar{S}_i . For $c\theta = 1$ and $c \ll 1$, we have

$$\frac{\Gamma(c + |\bar{S}_i|/2)}{(2\pi)^{|\bar{S}_i|/2} \theta^c \Gamma(c) \left(\frac{1}{2} \sum_{k \in \bar{S}_i} |y_k|^2 + \frac{1}{\theta}\right)^{c+|\bar{S}_i|/2}} \propto \frac{1}{\left(\sum_{k \in \bar{S}_i} |y_k|^2\right)^{|\bar{S}_i|/2}}.$$

The dependence on $(\sigma^2 + \rho^2) = 1/\beta$ still remains. This variance can be estimated for example by using the scheme described in Section 4.1.

We stress that with σ eliminated by marginalization, y_i become correlated, even if they were independent conditioned on σ . Hence the detection problem does not decouple, and we must compute (1). This involves a summation of $O(2^M)$ terms. In what follows we propose a way of dealing with this.

5. DETECTOR APPROXIMATIONS

Generally the optimal detector consists of computing (2), which contains 2^M terms. This must be done for each of the M observations. For large M this computation will be very burdensome. Only if σ, ρ are known, or considered known (by previous estimation), so that y_i become independent, (2) simplifies into (4). Hence, we have to approximate the sum in (2).

To approximate (2) we propose to use an algorithm presented in [7]. The idea is, that instead of considering all possible hypotheses $\{0, \dots, 2^M - 1\}$, we only consider a subset \mathcal{H} of them for which $P(H_k|y)$ is significant. We also have to normalize $P(H_k|y)$ for all $k \in \mathcal{H}$ so that they sum up to one. The probability of the event Ω_i , that a signal is present in observation i , is thus approximated by

$$P(\Omega_i|y) \approx \frac{1}{\sum_{m \in \mathcal{H}} P(y|H_m)P(H_m)} \sum_{k \in \mathcal{H}: i \in S_k} P(y|H_k)P(H_k),$$

That is, we sum over all hypotheses in \mathcal{H} which are likely to contain a signal in the i th observation. This yields the following equivalent

soft decision value

$$\lambda_i = \log \left(\frac{\sum_{k \in \mathcal{H}: i \in S_k} P(y|H_k)P(H_k)}{\sum_{k \in \mathcal{H}: i \in \bar{S}_k} P(y|H_k)P(H_k)} \right). \quad (7)$$

The set \mathcal{H} of indices k for which $P(H_k|y)$ is significant is chosen as follows [7]:

1. Start with a set $\mathcal{B} = \{1, 2, \dots, M\}$ and a hypothesis H_i (H_0 or H_{2^m-1} are natural choices).
2. Compute the contribution to (7), $P(H_i|y)$.
3. Evaluate $P(H_k|y)$ for all H_k which can be obtained from H_i by changing the state of one observation y_j , $j \in \mathcal{B}$. That is, if $y_j = x_j + n_j$ in H_i , then $y_j = n_j$ in H_k and vice versa. Choose the j which yields the largest $P(H_k|y)$. Set $i := k$ and remove j from \mathcal{B} .
4. If $\mathcal{B} = \emptyset$ (this will happen after M iterations), compute the contribution of the last H_i to (7) and then terminate. Otherwise, go to Step 2.

This algorithm will change the state of each observation once, and choose the largest term from each level. The sums of (7) will finally contain $M + 1$ terms instead of 2^M .

6. NUMERICAL RESULTS

We show some numerical results for the cooperative spectrum sensing application. We considered 5 cooperating sensors that scan $M = 100$ bands. All results are obtained by Monte-Carlo simulation in a standard manner, and performance is given as the probability P_{MD} of a missed detection of Ω_i as function of probability of a false alarm P_{FA} . In all simulations the true parameter values were $\sigma^2 = 1$ for the noise variance, $\rho^2 = 36$ for the signal variance, and $p = 0.5$ for the probability of a signal presence in a given band.

Example 1: Comparison of Detectors (Figure 1). We first compare the following schemes:

- (i) Optimal detection, known variances: (2)–(3), using true σ, ρ .
- (ii) Optimal detection, estimated variances: (2)–(3) and (5)–(6)
- (iii) Approximation algorithm, known variances: algorithm of Section 5, using true σ, ρ .
- (iv) Approximation algorithm, σ^2 by marginalization, $\sigma^2 + \rho^2$ by estimation (see Section 4.1)

Throughout, we use the true value of p in (2). Figure 1 shows the results. We observe that the scheme with estimated variances (ii) performs better than the scheme (iv) with marginalized noise variance. One reason for this is that the detector based on estimation of σ uses more a priori information (for example, p is used explicitly in the estimation). In addition, the marginalization-based scheme uses the approximate algorithm of Section 5, whereas with estimated noise variance we can use (4).

Example 2: Sensitivity to errors in p (Figure 2). So far we have assumed that perfect knowledge of p was available. In this example we will examine how performance degrades when the a priori knowledge of p is imperfect. Figure 2 shows the result. In all simulations, $p = 0.5$ was used to generate the data. We note that for the estimation scheme, it seems to be better to overestimate than to underestimate p . Underestimation yields a small decrease in performance, whereas the performance with overestimation is almost as good as with perfect knowledge. For the marginalization scheme, the performance increases for large P_{fa} when p is underestimated.

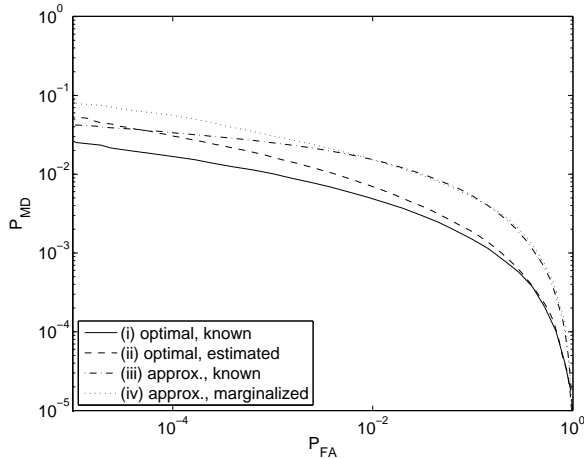


Fig. 1. ROC curves for the different detection schemes with cooperation among 5 sensors. In this example, $\sigma^2 = 1$, $\rho^2 = 36$, $p = 0.5$.

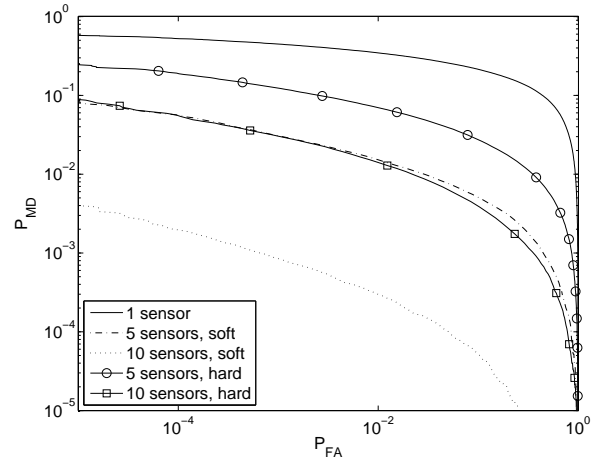


Fig. 3. ROC curves for different number of cooperating sensors. In this example $\sigma^2 = 1$, $\rho^2 = 36$, $p = 0.5$.

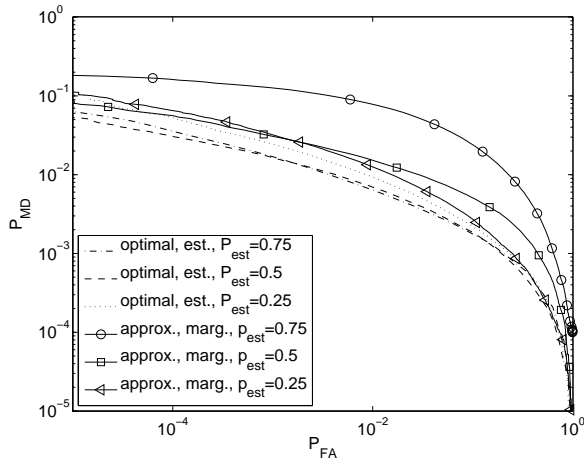


Fig. 2. ROC curves with imperfect knowledge of p . Data were generated using $\sigma^2 = 1$, $\rho^2 = 36$, and $p = 0.5$.

We believe the reason lies in the suboptimality of the approximation algorithm of Section 5.

Example 3: Cooperative spectrum sensing (Figure 3). We next illustrate the benefit of cooperation, and especially combination of soft decisions. For this example we use the detection scheme (iv) above (marginalized noise variance, approximate detector). Similar results can be obtained for the other schemes. We also compare with the case where the sensors only transmit binary values (hard decisions) for each band to the fusion center (this is equivalent to quantizing $\lambda_{c,i}$ to ± 1). Figure 3 shows the results of these simulations. We see the large gains of cooperation, and the gains of using soft information.

7. CONCLUDING REMARKS

We have dealt with a fundamental problem that has applications in many areas, multiband spectrum sensing being the most important driving motivator for our work. The difficulty of the problem lies in the fact that, on the one hand one would prefer a detector that makes

no a priori assumptions. On the other hand, without any prior knowledge at all the problem does not seem well defined (at least in a pure Bayesian framework), and we had to proceed by inserting estimated parameter values into the formal expressions for the posterior probabilities.

We modeled both signal and noise as zero-mean Gaussian variables. This is a fairly simple model, but it allowed us to expose the fundamental difficulties with the unknown noise variance. Future work may include extensions of the signal model, for example to work with feature vectors instead of scalar observations.

8. REFERENCES

- [1] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Wideband spectrum sensing in cognitive radio networks," *Proc. of IEEE ICC*, pp. 901–906, May 2008.
- [2] A. Taherpour, S. Gazor, and M. Nasiri-Kenari, "Wideband spectrum sensing in unknown white Gaussian noise," *IET Communications*, vol. 2, no. 6, pp. 763–771, July 2008.
- [3] S. M. Mishra, A. Sahai, and R. W. Brodersen, "Cooperative sensing among cognitive radios," in *Proc. of IEEE ICC*, vol. 4, pp. 1658–1663, June 2006.
- [4] J. Ma and Y. Li, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," *Proc. of IEEE GLOBECOM*, pp. 3139–3143, Nov. 2007.
- [5] E. Gudmundson and P. Stoica, "On denoising via penalized least-squares rules," *Proc. of IEEE ICASSP*, pp. 3705–3708, March 2008.
- [6] L. Wei, N. Kumar, V. N. Lolla, E. Keogh, S. Lonardi and C. A. Ratanamahatana, "Assumption-free anomaly detection in time series", *Proc. of SSDBM*, June 2005.
- [7] E. G. Larsson and Y. Selén, "Linear regression with a sparse parameter vector," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 451–460, Feb. 2007.
- [8] D. J. C. Mackay, *Information Theory, Inference & Learning Algorithms*, Cambridge University Press, June 2002.