IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE.Please read the form carefully and keep a copy for your files.

Author Online Use

6. Personal Servers. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore.Authors shall not post the final, published versions of their papers.

7. Classroom or Internal Training Use. An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the authors personal web site or the servers of the authors institution or company in connection with the authors teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.

HYPOTHESIS TESTING BY USING QUANTIZED OBSERVATIONS

Cathel Zitzmann, Rémi Cogranne, Florent Retraint, Igor Nikiforov, Lionel Fillatre, Philippe Cornu

ICD - LM2S - Université de Technologie de Troyes - UMR STMR CNRS
12, rue Marie Curie - B.P. 2060 - 10010 Troyes cedex - France
E-mail : firstname.lastsname@utt.fr

ABSTRACT

The goal of this paper is to study the hypothesis testing using a parametric statistical model with nuisance parameters based on quantized observations and related to the detection of hidden information.

Index Terms— Statistical decision, quantized observations, parametric model, asymptotic local approach, hidden information detection.

1. INTRODUCTION AND CONTRIBUTION

In the last two decades substantial progress has been made in the detection of hidden information or hidden communication channels in media files or streams. Typically, it is necessary to reliably detect in a huge set of files (image, audio, and video) which of these files contain the hidden information (like a text, an image,..). An important challenge is to get the hidden information detection algorithms with analytically predictable probabilities of false alarm and missed detection. The following theoretical problems remain unsolved :

- How to deal with the quantized observations? How does the quantization impact the probabilities of false alarm and missed detection ?
- What is the benefits from using a parametric statistical model of cover media (or cover channel) for hidden information detection ?

Hence, the goal of this paper is to study the theoretical aspects of hypothesis testing using a parametric statistical model with nuisance parameters based on quantized observations and related to the detection of hidden information. It is worth noting that the existence of a quantizer between the sensor and the estimation algorithm leads to the increasing complexity of estimation methods. Many results from the classical estimation theory are not applicable to quantized data (for example, the Gauss-Markov theorem). Some results on the estimation by using quantized observations are available in the literature (see for instance [1, 2]). In contrast to the local likelihood ratio quantization in distributed detection (see for instance [3]), dealing with quantized observations in the presence of nuisance parameters in hypothesis testing is almost unknown.

The contribution of the paper is threefold. First, the impact of the quantization on the probability of false alarm and missed detection is studied. Equations for first two moments of the log likelihood ratio are obtained. An asymptotic expression of the test power as a function of the false alarm rate is given by Theorem 1. Second, when the embedding rate is unknown, an asymptotic local uniformly most powerful test is designed. Third, a realistic (regressive) model of cover media is integrated in the statistical GLR-type test. This test is almost invariant. It is shown that this test is closely related to the WS steganalysers reputed as very efficient [4].

The paper is organised as follows. Section 2 is devoted to the case of perfectly known statistical model of cover media. The embedding rate is also assumed to be known. Section 3 considers the case of unknown embedding rate and a more realistic model of cover media. The theoretical comparison between the proposed and some heuristic steganalysers is discussed here. Finally, some conclusions are drawn in Section 4.

2. STATISTICAL DECISION BASED ON QUANTIZED OBSERVATIONS

2.1. Model of quantized cover media

Let us assume that the observation vector $C_n = (c_1, \dots, c_n)^T$ which characterizes a cover media is defined in the following manner :

$$C_n = Q_1[Y_n], \quad Y_n \sim P_\theta, \tag{1}$$

where $Q_1[y_i] = \lfloor y_i \rfloor$ is the operation of uniform quantization (integer part of y_i) and the vector $Y_n = (y_1, \dots, y_n)^T$ follows the distribution P_θ parameterized by the parametric vector θ which describes the properties of media files or streams. In the framework of hidden information detection, θ is a nuisance parameter. The binary representation of c (the index is omitted to seek simplicity) is $c = Q_1[y] = \sum_{i=0}^{q-1} b_i 2^i$, where $b_i = \{0, 1\}$. A simplified model of quantization is used in this paper. It is assumed that the saturation is absent, i.e. the

This work is supported by French National Agency (ANR) through ANR-CSOSG Program (Project ANR-07-SECU-004)

probability of the excess over the boundary 0 and 2^q-1 for the observation y is negligible.

2.2. Problem statement : test between two hypotheses

First, let us define two alternative hypotheses for one quantized observation z (seeking simplicity) :

$$\mathcal{H}_0 \ : \ z=c=Q_1[y] \sim Q_{Q_1}=[q_0,\ldots,q_{2^q-1}]$$

and

$$\mathcal{H}_1 \ : z=\begin{cases} Q_2[y]+z_s & \text{with probability} \quad R \\ c=Q_1[y] & \text{with probability} \quad 1-R, \end{cases}$$

where R is the embedding rate, $Q_2[y]=\sum_{i=1}^{q-1} b_i 2^i$, is an uniform quantization by using 2^{q-1} thresholds, $Q_2[y] \sim Q_{Q_2}$, $z_s \sim Q_s = \mathrm{B}(1,p)$ is the Bernoulli distribution which defines the hidden information (usually $p=0.5$). To get the double quantization $Q_2[y]$ from $Q_1[y]$ it is assumed that $b_0 \equiv 0$. Under hypothesis $\mathcal{H}_1$, the Least Significant Bit $\mathrm{LSB}(Q_1[y])=b_0$ is used as a container of hidden information because the LSB-based steganography provides serious embedding capacity without introducing significant distortions. The discrete distributions $Q_{Q_1}(\ldots)$, $Q_{Q_2}(\ldots)$ represent the quantized observation z without hidden information and Q_s represents the hidden information.

2.3. A known embedding rate. Two simple hypotheses : likelihood ratio test

Let us suppose that the distributions Q_s, Q_{Q_1}, Q_{Q_2} and the embedding rate R are exactly known. In this case the likelihood ratio (LR) for one observation is written as follows :

$$\Lambda_R(z)=R\frac{Q_{Q_2}\left(Q_2[y]\right)}{2Q_{Q_1}\left(Q_1[y]\right)}+(1-R) \tag{2}$$

The most powerful (MP) Neyman-Pearson test δ over the class

$$\mathcal{K}_{\alpha_0}=\left\{\delta : \mathbb{P}_0(\delta(Z_n)=\mathcal{H}_1) \le \alpha_0\right\},$$

where $\mathbb{P}_i(\ldots)$ denotes the probability under hypothesis $\mathcal{H}_i$, $i=0,1$, is given by the following decision rule :

$$\delta_R(Z_n)=\begin{cases} \mathcal{H}_0 \ \text{ if } \ \Lambda_R(Z_n)=\prod_{i=1}^{n}\Lambda_R(z_i) \ < \ h \\ \mathcal{H}_1 \ \text{ if } \ \Lambda_R(Z_n)=\prod_{i=1}^{n}\Lambda_R(z_i) \ \ge \ h \end{cases}. \tag{3}$$

The threshold h is defined as a solution of $\mathbb{P}_0(\Lambda_R(Z_n) \ge h)=\alpha_0$. The MP test $\delta_R(Z_n)$ maximizes the power

$$\beta_{\delta_R}=1-\mathbb{P}_1(\delta_R(Z_n)=\mathcal{H}_0)=1-\alpha_1$$

over the class $\mathcal{K}_{\alpha_0}$.

2.4. The moments of approximate log likelihood ratio

Let $Y_n \sim \mathcal{N}(\theta,\sigma^2)$. The approximation of the log LR $\log\Lambda_R(Z_n)$ (see equation (2)) by neglecting the corrective term related to quantized Gaussian law and under assumption that $R=1$ is given by

$$\log\widetilde{\Lambda}_1(Z_n)=\sum_{i=1}^{n}\frac{1}{2\sigma^2}\left[-\left(Q_2[y_i]+1-\theta\right)^2+\left(Q_1[y_i]+0.5-\theta\right)^2\right]. \tag{4}$$

It follows from the central limit theorem that the fraction

$$\frac{\log\widetilde{\Lambda}_1(Z_n)-n\mathbb{E}_i\left(\log\widetilde{\Lambda}_1(z)\right)}{\sigma_i\sqrt{n}} \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0,1), \quad i=0,1,$$

where $\sigma_i^2=\mathrm{Var}_i\left(\log\widetilde{\Lambda}_1(z)\right)$, will converge in distribution to the standard normal distribution as n goes to infinity. The expectation and variance are denoted by $\mathbb{E}_i(\ldots)$ and $\mathrm{Var}_i(\ldots)$ under $\mathcal{H}_i$, respectively. Hence, to compute the error probabilities it is necessary to get the expectations and variances of the approximate log LR. Under hypothesis $\mathcal{H}_0$, the expectation of the approximate log LR is given by the following expression

$$m_0=\mathbb{E}_0\left[\log\widetilde{\Lambda}_1(z)\right]=-\frac{1}{8\sigma^2}+\frac{\varepsilon}{\sigma^2}, \tag{5}$$

where the coefficient ε defines the impact of the quantization. It can be proved that this coefficient is given by

$$\begin{aligned}\varepsilon &\overset{\text{def.}}{=} \mathbb{E}_0\left[\zeta(b_0-0.5)\right] \\ &=\sum_{m=-\infty}^{\infty}\left[\Phi\left(\frac{2m+2-\theta}{\sigma}\right)-\Phi\left(\frac{2m+1-\theta}{\sigma}\right)\right]\frac{(2m+1.5-\theta)}{2} \\ &-\sum_{m=-\infty}^{\infty}\left[\Phi\left(\frac{2m+1-\theta}{\sigma}\right)-\Phi\left(\frac{2m-\theta}{\sigma}\right)\right]\frac{(2m+0.5-\theta)}{2}, \end{aligned} \tag{6}$$

where $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{u^2}{2}}du$, $\zeta_i=Q_1[y_i]+0.5-\theta$, $b_{0,i}=\mathrm{LSB}(Q_1[y_i])$. It can be also proved by analogy with the previous equation that

$$\sigma_0^2=\mathrm{Var}_0\left[\log\widetilde{\Lambda}_1(z)\right]=\frac{\mathbb{E}_0\left[\zeta^2\right]-4\varepsilon^2}{4\sigma^4}, \tag{7}$$

where

$$\mathbb{E}_0\left[\zeta^2\right]=\sum_{m=-\infty}^{\infty}\left[\Phi\left(\frac{m+1-\theta}{\sigma}\right)-\Phi\left(\frac{m-\theta}{\sigma}\right)\right](m+0.5-\theta)^2.$$

Under hypothesis $\mathcal{H}_1$, it is assumed that $b_{0,i}=z_{s,i}$. The expectation and variance of the approximate log LR are given by

$$m_1=\mathbb{E}_1\left[\log\widetilde{\Lambda}_1(z)\right]=\frac{1}{8\sigma^2}, \tag{8}$$

$$\sigma_1^2=\mathrm{Var}_1\left[\log\widetilde{\Lambda}_1(z)\right]=\frac{1}{4\sigma^4}\mathbb{E}_1\left[\xi^2\right], \tag{9}$$

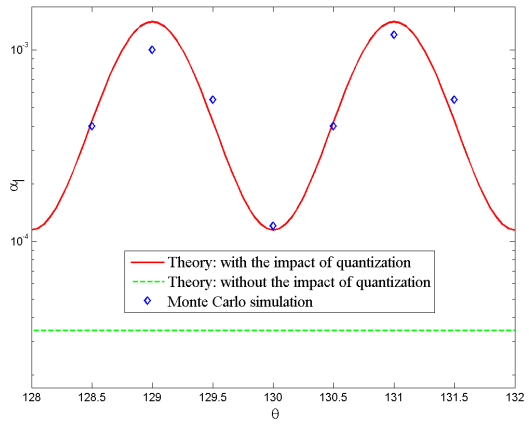

Fig. 1. The impact of the quantization on the probability of missed detection α_1.

where $\xi_i = Q_2[y_i] + 1 - \theta$. To illustrate the impact of the quantization, let us assume that : $\theta \in [128; 132]$, $\sigma = 1$ and $n = 200$. The required probability of false alarm is $\alpha_0 = 10^{-3}$. The comparison of theoretical equations for α_1 with the Monte Carlo simulation (10^6 repetitions) is presented in Figure 1. The impact of quantization on the probability of missed detection α_1 is significant.

Theorem 1 *Let the true embedding rate be* $\widetilde{R} : 0 < \widetilde{R} \leq 1$ *but the log LR (4) is computed under assumption that* $R = 1$. *The power of this test with taking into account the impact of quantization is approximately given by (for large* n*) :*

$$\beta_{\delta_1} \simeq 1 - \Phi\left(\Phi^{-1}(1-\alpha_0)\frac{\sigma_0}{\sigma_{\widetilde{R}}} - \frac{(m_1-m_0)\widetilde{R}\sqrt{n}}{\sigma_{\widetilde{R}}}\right) \quad (10)$$

where $\sigma_{\widetilde{R}} = f(m_0, m_1, \sigma_0, \sigma_1)$, m_i *and* σ_i *are computed by using equations (5) - (9).*

3. AN UNKNOWN EMBEDDING RATE

3.1. Two composite hypotheses

Let us assume that the distributions Q_s, Q_{Q_1}, Q_{Q_2} are known, but the embedding rate R is unknown. The following alternative composite hypotheses have to be tested by using n observations Z_n representing the cover media :

$$\mathcal{H}_0 = \{R \leq r^*\} \text{ against } \mathcal{H}_1 = \{R > r^*\} \quad (11)$$

Hence, the LR (2) becomes

$$\Lambda_{R_0,R_1}(Z_n) = \prod_{i=1}^{n} \frac{R_1 \frac{1}{2} Q_{Q_2}(Q_2[y_i]) + (1-R_1) Q_{Q_1}(Q_1[y_i])}{R_0 \frac{1}{2} Q_{Q_2}(Q_2[y_i]) + (1-R_0) Q_{Q_1}(Q_1[y_i])}, \quad (12)$$

where $R_0 \leq r^* < R_1$. The main difficulty is that the values of acceptable R_0 and unacceptable R_1 embedding rates are unknown. The ultimate challenge for anyone in the case of two composite hypotheses is to get a uniformly MP (UMP) test δ which maximizes the power function $\beta(R) = 1 - \mathbb{P}_R(\delta(Z_n) = \mathcal{H}_0)$, where $\mathbb{P}_R(\ldots)$ denotes the probability under the assumption that the embedding rate is equal to R, for any $R > r^*$ over the class $\mathcal{K}_{\alpha_0} = \left\{\delta : \sup_{R \leq r^*} \mathbb{P}_R(\delta(Z_n) = \mathcal{H}_1) \leq \alpha_0\right\}$. An efficient solution is based on the asymptotic local approach proposed by L. Le Cam [5]. The idea of this approach is that the "distance" between alternative hypotheses depends on the sample size n in such a way that the two hypotheses get closer to each other when n tends to infinity. By using an asymptotic expansion of the log LR, a particular hypothesis testing problem can be locally reduced to a relatively simple UMP hypothesis testing problem between two Gaussian scalar means [5]. The log LR can be re-written by using the following asymptotic expansion

$$\log \Lambda\left(Z_n; \frac{1}{\sqrt{n}}\delta_r\right) \simeq \frac{1}{\sqrt{n}}\delta_r \zeta_n(Z_n; r^*) - \frac{1}{2}\delta_r^2 \mathcal{F}(r^*)$$

where $\mathcal{F}(R)$ is the Fisher information and the efficient score is given by

$$\zeta_n(Z_n; r^*) = \sum_{i=1}^{n} \zeta(z_i; r^*) = \sum_{i=1}^{n} \frac{\Lambda_1(z_i) - 1}{r^* \Lambda_1(z_i) + (1 - r^*)} \quad (13)$$

Therefore, the local UMP test to chose between two alternative hypotheses (11) is given by the following rule :

$$\delta_{r^*}(Z_n) = \begin{cases} \mathcal{H}_0 & \text{if } \zeta_n(Z_n; r^*) < h \\ \mathcal{H}_1 & \text{if } \zeta_n(Z_n; r^*) \geq h \end{cases},$$

where h is a solution of $\sup_{R \leq r^*} \mathbb{P}_R(\zeta_n(Z_n; r^*) \geq h) = \alpha_0$.

3.2. A more realistic model of cover media

As it follows from equation (10), the power β of an optimal test depends on the standard deviation σ of cover media for a given rate of false alarm α_0. Hence, to increase the power β, someone has to reduce the standard deviation σ by using a parametric model of cover media. Often the observation vector $C_n = (c_1, \ldots, c_n)^T$ which characterizes the cover media can be defined by the following regression model :

$$C_n = Q_1[Y_n], \quad Y_n = Hx + \xi \sim \mathcal{N}(Hx, \sigma^2 I_n)$$

where H is a known $[n \times l]$ full rank matrix, $n > l$, $x \in \mathbb{R}^l$ is a nuisance parameter and σ^2 is the residual variance. The vector C_n (pixels) is extracted from the cover media file (digital image, for instance) by using a specially chosen segment or mask. Such a parametric model is an efficient method to reduce the standard deviation σ. The new hypothesis testing problem with a parametric model of cover media consists in deciding between the following hypotheses

$$\mathcal{H}_0 : Z_n = C_n = Q_1[Y_n], \quad (14)$$

$$\mathcal{H}_1\colon z_i = \begin{cases} Q_2[y_i] + z_{s,i} & \text{with probability} \quad R \\ c_i = Q_1[y_i] & \text{with probability} \quad 1\text{–}R \end{cases}, \quad (15)$$

where $Y_n = (y_1, \ldots, y_n)^T \sim \mathcal{N}(Hx, \sigma^2 I_n)$. In practice, x and σ^2 are unknown. The theoretical aspects of dealing with nuisance parameters in the framework of statistical decision theory is discussed in [6]. An efficient approach to this problem is based on the theory of invariance in statistics. The optimal invariant tests and their properties in the context of image processing have been designed and studied in [7, 8]. The parameter vector x can be estimated by using $Q_2[Y_n]$ which is free from the embedding information. The "approximate" log GLR is given by

$$\log \widehat{\Lambda}_1(Z_n) \simeq \frac{1}{\widehat{\sigma}^2} \left[P_H Q_2[Y_n]\right]^T \left[B_0 - 0.5 \cdot \mathbf{1}_n\right] + \frac{n}{8\widehat{\sigma}^2}, \quad (16)$$

where $\widehat{\sigma}^2$ is the maximum likelihood estimation of σ^2 based on $Q_2[Y_n]$, I_n is an $(n \times n)$ identity matrix, $P_H = I_n - H(H^T H)^{-1} H^T$ is a projection matrix, $B_0 = (b_{0,1}, \ldots, b_{0,n})^T$ and $\mathbf{1}_n = (1, \ldots, 1)^T$. The idea of the invariant hypotheses testing approach is based on the existence of the natural invariance of the detection problem with respect to a certain group of transformation. Let us note that the above mentioned hypotheses testing problem given by (14) - (15) remains "almost" invariant under the group of translations $G = \{g : g(Y_n) = Y_n + Hx\}$, $x \in \mathbb{R}^l$. The word "almost" is due to the quantization $Q_j[y]$, $j = 1, 2$. Without the quantization, the invariance will be exact.

3.3. Relation between the proposed and some known heuristic tests

The first right hand side term in equation (16) defines the sensitivity of the test because the second right hand side term $\frac{n}{8\widehat{\sigma}^2}$ does not depend on the embedded secret message. The first right hand side term in equation (16) represents an inner product of the vector of "residuals" $\varepsilon = P_H Q_2[Y_n]$, i.e. the vector of the projection of Y_n on the orthogonal complement of the column space of H, and the vector $[B_0 - 0.5 \cdot \mathbf{1}_n]$ composed of $\mathrm{LSB}(Q_1[y_i]) - 0.5$:

$$\sum_{i=1}^{n} \overbrace{\widehat{\sigma}^{-2}}^{=\text{"weight"}} \cdot \overbrace{(Q_2[y_i] - (H\widehat{x})_i + 1)}^{=\text{"residual" } \varepsilon_i} \cdot \overbrace{(b_{0,i} - 0.5)}^{=\mathrm{LSB}(Q_1[y_i]) - 0.5} \quad . \quad (17)$$

Let us now compare the last equation with the recently developed WS steganalysers reputed very efficient [4]. These steganalysers are based on the following statistics :

$$\sum_{i=1}^{n} \overbrace{w_i}^{=\text{"weight"}} \cdot \overbrace{(z_i - \mathcal{F}(z)_i)}^{=\text{"residual" } \varepsilon_i} \cdot \overbrace{(z_i - \overline{z}_i)}^{=2 \cdot (\mathrm{LSB}(Q_1[y_i]) - 0.5)} \quad , \quad (18)$$

where $\mathcal{F}(s)$ denotes a "filter" dedicated to estimate the cover-image by filtering the stego-image, the weight w_i is chosen as $\frac{1}{1+\sigma_i^2}$, σ_i^2 is the "local" variance and $\overline{z}_i$ denotes the non-negative integer z_i with the LSB flipped. As it follows from equations (17) - (18), the steganalysers developed in [4] coincide with the first term of the tractable log GLR (16). Nevertheless, the second right hand side term $\frac{n}{8\widehat{\sigma}^2}$ of (16) is also necessary to correctly calculate the threshold h as a solution of the following equation

$$\mathbb{P}_0(\log \widehat{\Lambda}_1(Z_n) \geq h) = \alpha_0.$$

4. CONCLUSIONS

The impact of data quantization on the probability of false decision in the problem of hidden information detection has been studied. A local UMP test has been designed for the case of unknown embedding rate. An almost invariant test has been developed for a realistic model of cover media. A relation between this test and some known heuristic steganalysers has been established.

5. REFERENCES

[1] Le Yi Wang and G. George Yin, "Asymptotically efficient parameter estimation using quantized output observations," *Automatica*, vol. 43, pp. 1178–1191, July 2007.

[2] Fredrik Gustafsson and Rickard Karlsson, "Statistical results for system identification based on quantized observations," *Automatica*, vol. 45, pp. 2794–2801, December 2009.

[3] Rick S Blum, Saleem A Kassam, and H Vincent Poor, "Distributed detection with multiple sensors II. advanced topics," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 64–79, 1997.

[4] Andrew D. Ker and Rainer Bhme, "Revisiting weighted stego-image steganalysis," in *Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, volume 6819*. 2008, pp. 5 1 – 5 17, San Jose, CA, 27-31 January.

[5] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Series in Statistics, Springer, New York, 1986.

[6] E. L. Lehmann, *Testing Statistical Hypotheses*, 2nd edn. Springer, New York, 1986.

[7] Lionel Fillatre, Igor Nikiforov, and Florent Retraint, "ε-optimal non-bayesian anomaly detection for parametric tomography," *IEEE Transactions on Image Processing*, vol. 17, no. 11, pp. 1985–1999, 2008.

[8] L.L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2146–2157, 1994.