Modeling and Extending the Ensemble Classifier for Steganalysis of Digital Images Using Hypothesis Testing Theory

Rémi Cogranne, Jessica Fridrich

Modeling and Extending the Ensemble Classifier for Steganalysis of Digital Images Using Hypothesis Testing Theory

Rémi Cogranne, *Member, IEEE,* and Jessica Fridrich, *Senior Member, IEEE,*

Accepted version, final version available online on ieeexplore.ieee.org. DOI: 10.1109/TIFS.2015.2470220

***Abstract*—The machine learning paradigm currently predominantly used for steganalysis of digital images works on the principle of fusing the decisions of many weak base learners. In this paper, we employ a statistical model of such an ensemble and replace the majority voting rule with a likelihood ratio test. This allows us to train the ensemble to guarantee desired statistical properties, such as the false-alarm probability and the detection power while preserving the high detection accuracy of original ensemble classifier. It also turns out the proposed test is linear. Moreover, by replacing the conventional total probability of error with an alternative criterion of optimality, the ensemble can be extended to detect messages of an unknown length to address composite hypotheses. Finally, the proposed well-founded statistical formulation allows us to extend the ensemble to multi-class classification with an appropriate criterion of optimality and an optimal associated decision rule. This is useful when a digital image is tested for presence of secret data hidden by more than one steganographic method. Numerical results on real images show the sharpness of the theoretically established results and the relevance of the proposed methodology.**

***Index Terms*—Hypothesis testing theory, information hiding, optimal detection, multi-class classification, ensemble classifier.**

I. Introduction

STEGANOGRAPHY is often referred to as the science of covert communication. Its objective is to hide a secret message within an innocuous looking cover object creating thus a stego-object that can be sent over an insecure channel without raising any suspicion. Steganography that hides messages in digital images has received great attention since mid 1990's.

Accepted version, final version available online on ieeexplore.ieee.org. DOI: 10.1109/TIFS.2015.2470220

This work was supported by Air Force Office of Scientific Research under the research grant number FA9950-12-1-0124. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of AFOSR or the U.S. Government.

Dr. Rémi Cogranne is with the Lab. for System Modelling and Dependability, ICD, UMR 6281 CNRS, Troyes University of Technology, Troyes, France. This work has been done while Rémi Cogranne was a visiting scholar at Binghamton University. Email : remi.cogranne@utt.fr

Prof. Jessica Fridrich is with the Department of Electrical and Computer Engineering, Binghamton University, NY, 13902, USA. Email: fridrich@binghamton.edu

A. Modern Steganalysis

While steganography has been advanced to hide data more efficiently and more securely, see e.g. [1], the related field of steganography detection, steganalysis, has also been developing at a fast rate. Currently, there exist two main trends in steganalysis:

1) Optimal detectors, as referred to in [2], find an optimal statistical test with respect to a given criterion based on a statistical model of cover objects. One of the first optimal detectors was proposed for Least Significant Bit (LSB) replacement in [3] using a simple statistical model of pixels. The detection has been further improved by considering more sophisticated models [4]–[7]. The methodology of optimal detection has also been applied to LSB matching in the spatial domain [8] and to the Jsteg algorithm in the JPEG domain [9], [10].
2) A large portion of steganalysis methods today are implemented using machine learning. First, a feature representation of cover objects, that could reveal steganography, is selected. Then, a classifier is trained to distinguish between cover and stego features. One of the first feature based steganalysis can be found in [11], where the Fisher Linear Discriminant (FLD) was used for machine learning. Soon, more efficient machine learning techniques were proposed, such as the Support Vector Machines (SVM) [12], [13], combined with a plethora of representations of digital images. Recently, the FLD-based ensemble classifier [14] has been successfully introduced as a scalable alternative to SVMs for features of a high dimension.

Machine learning based steganalysis methods are usually much more powerful than optimal detectors derived from simple models. They can also be extended to multi-class detection [15] and payload estimation [13]. Their drawback,

however, is that the theoretical statistical properties of such detectors are generally unknown: the false-alarm and correct-detection probabilities are always evaluated empirically from a large set of digital images. While optimal detectors derived from a cover model perform much worse in practice, since modeling such complex objects as digital images is challenging, these detectors offer certain indisputable advantages. First, one can theoretically find an explicit formula for the statistical performance of the optimal test and guarantee a false-alarm rate, which is crucial in practical application when searching for hidden data in large data sets. Second, such framework provides valuable insight into the problem of how the properties of digital images impact statistical detectability, see [16], [17] for an example of steganography designed to minimize the impact on a statistical image model.

B. Contribution and Organization of this Paper

The present paper leverages the advantages of both approaches. By employing an accurate statistical model for the base learners' projections in an ensemble classifier [14], it is possible to replace the majority voting detector with a likelihood ratio test. This allows us to (1) establish the statistical properties of the proposed optimal test, (2) use alternative criteria for the detector design (e.g., the Neyman–Pearson criterion instead of the frequently used Bayesian criterion), and (3) extend the classifier to multi-class detection while using optimality criteria established within the detection theory. The new framework, when used with the Bayesian criterion, preserves the high detection performance of the original ensemble classifier. To the best of our knowledge, this approach, which casts the ensemble base learners' within hypothesis testing theory, has never been studied before.

Although the present paper focuses on steganalysis with ensemble classifiers whose base learners are FLDs, the proposed approach is applicable to other instances of ensembles whose base learners are linear (e.g., linear SVMs).

We now highlight the main contributions of this paper with respect to the original formulation of the ensemble classifier for steganalysis [14]:

1) The statistical model of base learners' projections, whose accuracy is verified on real data sets, allows us to formally establish the statistical properties of the proposed test. For instance, one can compute the threshold that guarantees a desired false-alarm rate together with the highest power one can expect. This extension of the ensemble also allows using other measures of performance than the usual minimal total probability P_E and to draw the receiver operating characteristic (ROC) curve. This part has already been partially published in [18].
2) Moreover, with the proposed approach, we show that a statistical test can be designed for the ensemble so that it remains optimal (under a mild assumption) when the embedding payload / hidden message length is unknown.
3) Finally, the ensemble is extended to a multi-hypotheses detection/classification problem. This is possible because we casted the ensemble classifier within the hypothesis testing theory based on the proposed statistical model. A novel criterion of optimality is proposed (and the associated optimal statistical test is presented) to preserve the constraint on the false-alarm probability that, up to our knowledge, has never been studied in steganalysis.

This paper is an extension of our prior work [18]. It provides three majors advancements with respect to this prior art: 1) a detailed description of the implementation, which has been modified to improve the accuracy; 2) extension to the case of an unknown payload together with a proof of optimality of the proposed test; and 3) extension to the multi-hypotheses (multiclass steganalysis) with a novel criterion of optimality.

Though the present paper focuses on steganalysis, the proposed methodology is general enough to be used as a supervised learning method in a broader context, such as for content retrieval [19], face recognition [20], and automatic annotation [21], which also employ ensemble classifiers.

The paper is organized as follows. Section II provides a brief summary of ensemble classifiers for steganalysis using high-dimensional feature-spaces. Then, Section III states the steganalysis problem within the framework of the statistical hypothesis testing theory. This is followed by a description of the proposed optimal Likelihood Ratio Test (LRT) including the study of its statistical properties. Section IV extends the proposed optimal detection approach to the case of multi-class classification through an original criterion of optimality never used in steganalysis. Numerical results on a large image database for both spatial and JPEG domain steganographic methods presented in Section V show the sharpness of the theoretical results and the relevance of the proposed methodology. Finally, Section VI summarizes the present work and concludes the paper.

II. FLD Ensemble Classifier (Background)

In the whole paper, matrices are represented with capital bold letters $\mathbf{X}$, vectors are denoted with lower case bold letters $\mathbf{x}$, scalars with lower case letters x, and sets and probability distributions with calligraphic capital letters $\mathcal{X}$.

With the rapidly increasing dimension of feature spaces for steganalysis, ensemble classifiers have received a great interest because their computational complexity scales favorably with respect to the feature dimensionality. However, their theoretical performance remains largely unstudied. The present paper focuses on the ensemble classifier as originally designed for the BOSS competition [22] and further developed in [14]. Each base learner is a Fisher Linear Discriminant classifier trained on a uniformly randomly selected subset of features.

Formally, let $\mathbf{f} \in \mathbb{R}^d$ be a (column) vector of d features extracted from one image. It is assumed that the features extracted from N pairs of cover and corresponding stego images, respectively denoted $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_N)$ and $\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_N)$, are available together with their class labels. This set is divided into two disjoint subsets of N^{trn} training and N^{tst} testing samples, $N^{\mathrm{trn}} + N^{\mathrm{tst}} = N$. During the training phase, each base learner is given a subset of features on which

an FLD classifier is trained.[1] Let the training sets of cover and stego image features be matrices of size $d \times N^{\mathrm{trn}}$ denoted $\mathbf{C}^{\mathrm{trn}} = (\mathbf{c}_1^{\mathrm{trn}}, \dots, \mathbf{c}_{N^{\mathrm{trn}}}^{\mathrm{trn}})$ and $\mathbf{S}^{\mathrm{trn}} = (\mathbf{s}_1^{\mathrm{trn}}, \dots, \mathbf{s}_{N^{\mathrm{trn}}}^{\mathrm{trn}})$. The FLD assumes that, within these two classes, the features are i.i.d. with means $\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_s$, of size $d \times 1$, and covariance matrices $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_s$ of size $d \times d$. Among all linear decision rules defined by:[2]

$$\mathcal{C} : \begin{cases} \mathcal{H}_0 \text{ if } & \mathbf{w}^{\mathrm{T}}\mathbf{f} < b \\ \mathcal{H}_1 \text{ if } & \mathbf{w}^{\mathrm{T}}\mathbf{f} > b \end{cases} \tag{1}$$

where $\mathbf{f}$ is a feature vector to be classified, the FLD finds the weighting vector $\mathbf{w} \in \mathbb{R}^d$ that maximizes the following Fisher separability criterion:

$$\frac{\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_s)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_s)^{\mathrm{T}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_s)\mathbf{w}}. \tag{2}$$

Few calculations show that maximizing the Fisher criterion on the training data, $\mathbf{C}^{\mathrm{trn}}$ and $\mathbf{S}^{\mathrm{trn}}$, gives the following weighting vector $\mathbf{w}$:

$$\mathbf{w} = \left(\widehat{\boldsymbol{\Sigma}}_c + \widehat{\boldsymbol{\Sigma}}_s\right)^{-1} (\widehat{\boldsymbol{\mu}}_c - \widehat{\boldsymbol{\mu}}_s) \tag{3}$$

$$\text{with } \widehat{\boldsymbol{\mu}}_c = \frac{1}{N_{\mathrm{trn}}} \sum_{n=1}^{N^{\mathrm{trn}}} \mathbf{c}_n^{\mathrm{trn}}, \ \widehat{\boldsymbol{\mu}}_s = \frac{1}{N_{\mathrm{trn}}} \sum_{n=1}^{N^{\mathrm{trn}}} \mathbf{s}_n^{\mathrm{trn}}$$

$$\widehat{\boldsymbol{\Sigma}}_c = \frac{1}{N_{\mathrm{trn}} - 1}(\mathbf{C}^{\mathrm{trn}} - \widehat{\boldsymbol{\mu}}_c)(\mathbf{C}^{\mathrm{trn}} - \widehat{\boldsymbol{\mu}}_c)^{\mathrm{T}},$$

$$\text{and } \widehat{\boldsymbol{\Sigma}}_s = \frac{1}{N_{\mathrm{trn}} - 1}(\mathbf{S}^{\mathrm{trn}} - \widehat{\boldsymbol{\mu}}_s)(\mathbf{S}^{\mathrm{trn}} - \widehat{\boldsymbol{\mu}}_s)^{\mathrm{T}}.$$

The power of the ensemble classifier comes from using L different FLD classifiers all built on randomly selected subsets of d_{sub} features denoted $\mathcal{F}_1, \dots, \mathcal{F}_L$. It is worth noting that the vector $\mathbf{v} \in \mathbb{R}^L$ of all L projections from (1) can be written:

$$\mathbf{v} = \mathbf{P}\mathbf{f}, \tag{4}$$

where, again, $\mathbf{f} \in \mathbb{R}^d$ is a feature vector to be classified and $\mathbf{P}$ is a "sparse" matrix of size $L \times d$ whose l-th row contains zeros for all features not included in $\mathcal{F}_l$ while it contains the weighting vector from the corresponding l-th base learner in all remaining elements.

In the ensemble classifier as developed for steganalysis [14], the projection $\mathbf{v}$ is then thresholded, to obtain a vote from each base learner, as follows:

$$\operatorname{sign}(\mathbf{v} - \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{R}^L$ represents the vector of thresholds of all L base learners (1) and the $\operatorname{sign}(x)$ function, applied element-wise, is 1 if x is positive, -1 is x is negative and 0 in the (unlikely) case of $x = 0$.

In this paper, we model the distribution of $\mathbf{v}$ using the multivariate Gaussian distribution and cast the process of reaching the ensemble decision within the framework of hypothesis testing. This will allow us to design optimal detectors and establish their performance. In Section IV, we extend this approach to the case of multi-class steganalysis (multiple hypothesis testing). Because we also changed the implementation of the ensemble classifier for steganalysis, we provide a detailed comparison with the original ensemble in Section III-C.

[1] We also point out that in steganalysis the training set must always consist of pairs of cover-stego images as it has been shown that preserving those pairs can significantly increase the detection performance [23].

[2] Since the FLD is a well-known tool, it is only briefly described in this paper; the reader is referred to [24] for a more detailed exposition.

III. Optimal Binary Detector using Ensemble Classifier

Let us assume that the vector of base learners' projections $\mathbf{v}$, see Equation (4), follows the distribution $\mathcal{P}_{\theta_0}$ under the null hypothesis $\mathcal{H}_0$ (features are extracted from cover images) and $\mathcal{P}_{\theta_1}$ under the alternative hypothesis $\mathcal{H}_1$ (features extracted from stego-images with data hidden with a known payload R and a known embedding method). This constitutes the ideal scenario for a steganalyst as he/she knows the probability distribution under both hypotheses, the embedding method, and the payload R. Accepting for a moment this ideal setting, steganalysis amounts to choosing between the two following simple hypotheses:

$$\begin{cases} \mathcal{H}_0 : & \{\mathbf{v} \sim \mathcal{P}_{\theta_0}\}, \\ \mathcal{H}_1 : & \{\mathbf{v} \sim \mathcal{P}_{\theta_1}\}. \end{cases} \tag{5}$$

A statistical test is a mapping $\delta : \mathbb{R}^L \mapsto \{\mathcal{H}_0, \mathcal{H}_1\}$, such that hypothesis $\mathcal{H}_i$ is accepted if $\delta(\mathbf{v}) = \mathcal{H}_i$ (see [25], [26] for details). The present paper focuses on the Neyman–Pearson bi-criteria approach that aims at minimizing the missed-detection probability for a guaranteed false-alarm probability. Hence, let:

$$\mathcal{K}_{\alpha_0} = \{\delta : \mathbb{P}_{\mathcal{H}_0}(\delta(\mathbf{v}) = \mathcal{H}_1) \leq \alpha_0\}, \tag{6}$$

be the class of tests with a false-alarm probability upper-bounded by α_0. Here, $\mathbb{P}_{\mathcal{H}_i}(A)$ stands for the probability of event A under hypothesis $\mathcal{H}_i, i = \{0, 1\}$.

Among all tests in $\mathcal{K}_{\alpha_0}$, it is necessary to find a test δ that maximizes the power function defined by the correct detection probability:

$$\beta_\delta = \mathbb{P}_{\mathcal{H}_1}(\delta(\mathbf{v}) = \mathcal{H}_1), \tag{7}$$

which is equivalent to minimizing the missed-detection probability $\alpha_1(\delta) = 1 - \beta_\delta$.

When the hypotheses are simple, it follows from the Neyman–Pearson Lemma [26, Theorem 3.2.1] that the Most Powerful (MP) test in the class $\mathcal{K}_{\alpha_0}$ (6) is the Likelihood Ratio test (LRT) defined as:

$$\delta^{\mathrm{lr}}(\mathbf{v}) = \begin{cases} \mathcal{H}_0 & \text{if } \Lambda^{\mathrm{lr}}(\mathbf{v}) = \dfrac{p_{\theta_1}(\mathbf{v})}{p_{\theta_0}(\mathbf{v})} \leq \tau^{\mathrm{lr}}, \\ \mathcal{H}_1 & \text{if } \Lambda^{\mathrm{lr}}(\mathbf{v}) = \dfrac{p_{\theta_1}(\mathbf{v})}{p_{\theta_0}(\mathbf{v})} > \tau^{\mathrm{lr}}, \end{cases} \tag{8}$$

where p_{θ_0} and p_{θ_1} denote the joint probability density function (pdf) associated with the distributions $\mathcal{P}_{\theta_0}$ and $\mathcal{P}_{\theta_1}$, respectively, and τ^{lr} is the solution of the equation $\mathbb{P}_{\mathcal{H}_0}\left(\Lambda^{\mathrm{lr}}(\mathbf{v}) > \tau^{\mathrm{lr}}\right) = \alpha_0$ to ensure that the LRT is in the class $\mathcal{K}_{\alpha_0}$, see Equation (6).

The choice of the Neyman–Pearson criterion of optimality is justified by practical consideration; in fact, when analyzing a large number of digital images the most difficult challenge is to

guarantee a (very) low false-alarm probability, which is exactly the goal of the Neyman–Pearson approach that maximizes the detection accuracy under the constraint of a prescribed false-alarm probability. Other criteria, such as Fishers' separability criterion (2), Bayesian criterion, see Corollary 1, do not indicate how to set the threshold in order to guarantee a false-alarm probability and, consequently, do not provide an expression for the achievable detection accuracy for a given prescribed false-alarm rate.

The ideal scenario of testing simple hypotheses is addressed in the remainder of this section. Then, two extensions are presented within the same framework that aim at designing a detector with established statistical properties to guarantee a prescribed false-alarm probability. First, in Section III-E a practical application of the proposed test when the payload is unknown is addressed. Second, the problem of extending the proposed approach of optimal detection to multi-class classification is addressed in Section IV. This corresponds to classifying images with data hidden with different embedding methods.

A. Statistical Model of Ensemble Classifier

In the present paper, it is proposed to model the vector $\mathbf{v}$ of base learners' projections by a multivariate normal distribution. Fundamentally, it is hardly possible to formally prove that this model holds true whatever the features might be. However, since the number of features used by each base learner is usually quite large the use of the multivariate normal distribution is supported by invoking Lindeberg's central limit theorem (CLT) [26, Theorem 11.2.5]. This is later confirmed experimentally. Using this statistical model, $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, under the null hypothesis $\mathcal{H}_0$, and $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ under the alternative hypothesis $\mathcal{H}_1$. Here $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent the expectations and covariances of base learners' projections under hypotheses $\mathcal{H}_i$, $i = \{0, 1\}$.

In order to simplify the presentation of the proposed test, we transform the base learners' projections as follows:

$$\widetilde{\mathbf{v}} = \boldsymbol{\Sigma}_0^{-1/2} (\mathbf{v} - \boldsymbol{\mu}_0), \tag{9}$$

where the symbol $\boldsymbol{\Sigma}_0^{-1/2}$ denotes the symmetric matrix satisfying $\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{-1/2} = \boldsymbol{\Sigma}_0^{-1}$. It is straightforward to note that it is equivalent to computing the normalized projections $\widetilde{\mathbf{v}}$ from base learners projections $\mathbf{v}$ or from $\mathbf{v} - \mathbf{b}$, used to compute the votes, as the mean $\boldsymbol{\mu}_0$ is also modified in the same way. This is why, within the proposed methodology, there is no need to use any thresholds of the individual base learners. The affine transformation (9) guarantees that, under the hypothesis $\mathcal{H}_0$, the "normalized" base learners' projections $\widetilde{\mathbf{v}}$ follow a multivariate normal distribution with zero mean and identity covariance matrix: $\widetilde{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_L)$ with $\mathbf{I}_L$ the identity matrix of size L. It is important to note that the family of multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ remains invariant under such a transformation, see [25, Chap. 4] and [26, Chap. 6] for details about the invariance principle in statistical decision theory.

In this paper it is further assumed that the covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are equal; this assumption is approximately true for small payloads R, which are the focus of the present paper because it is the most difficult case for detection.

Let us denote

$$\boldsymbol{\theta_1} = \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{10}$$

The steganalysis detection problem can be rewritten as a choice between the two following simple hypotheses:

$$\begin{cases} \mathcal{H}_0 : & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_L)\}, \\ \mathcal{H}_1 : & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\theta}_1, \mathbf{I}_L)\}. \end{cases} \tag{11}$$

We have conducted a wide range of experiments to confirm the validity of this model. Some are presented among the numerical results in Section V-B.

B. Optimal LRT and Study of its Statistical Performance

As discussed in the introduction of Section III, for simple hypotheses (11) the optimal statistical test that guarantees a false-alarm probability and maximal power function is the LRT (8). In our case, the Likelihood Ratio (LR) between the tested hypotheses can be simplified as (see Appendix A):

$$\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) = \frac{\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}}}{\|\boldsymbol{\theta}_1\|}, \tag{12}$$

where, $\|\boldsymbol{\theta}_1\|^2 = \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_1$.

Note that the proposed LR, $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$, is related to the Linear Discriminant Analysis (LDA) because both share the underlying model, the homoscedastic multivariate Gaussian distribution of data, and reach the decision in a linear fashion based on the projection vector $\boldsymbol{\theta_1}$, see Eq. (10).
From the properties of the multivariate normal distribution, it immediately follows from the distribution of $\widetilde{\mathbf{v}}$ under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, see Equation (11), that the LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$, Equation (12), follows the following distribution:

$$\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) = \frac{\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}}}{\|\boldsymbol{\theta}_1\|} \sim \begin{cases} \mathcal{N}(0, 1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\|\boldsymbol{\theta}_1\|, 1) & \text{under } \mathcal{H}_1. \end{cases} \tag{13}$$

From here, it is straightforward to establish the statistical properties of the proposed LRT (8):

Proposition 1. *For any false alarm probability $\alpha_0 \in (0, 1)$, it follows from (13) that the following decision threshold:*

$$\tau^{\mathrm{lr}} = \Phi^{-1}(1 - \alpha_0), \tag{14}$$

guarantees that $\mathbb{P}_{\mathcal{H}_0}\left(\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) > \tau^{\mathrm{lr}}\right) = \alpha_0$. Here Φ and Φ^{-1} denote the normal cumulative distribution function (cdf) and its inverse.
From the expression for the threshold τ^{lr} (14), and the statistical distribution of the LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$, Equation (13), the power function of the most powerful LRT δ^{lr} is given by:

$$\begin{aligned} \beta_{\delta^{\mathrm{lr}}} = \mathbb{P}_{\mathcal{H}_1}\left(\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) > \tau^{\mathrm{lr}}\right) &= 1 - \Phi\left(\tau^{\mathrm{lr}} - \|\boldsymbol{\theta}_1\|\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \|\boldsymbol{\theta}_1\|\right). \end{aligned} \tag{15}$$

Proof. Proposition 1 is proved in Appendix A. □

Two essential elements can be deduced from Proposition 1. First, thanks to the normalization of the base learners' projections, see (9), and of the LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$ through the multiplication

by $\|\boldsymbol{\theta}_1\|^{-1}$ (12), the decision threshold only depends on the prescribed false-alarm probability. Second, the performance of the proposed optimal LRT is entirely given by $\|\boldsymbol{\theta}_1\|$, the norm of the expectation under $\mathcal{H}_1$.

Remark 1. *Note that the proposed LRT is in fact a linear classifier. It is indeed straightforward from Equation (4) and (9) that for a feature vector* $\mathbf{f}$ *the proposed LRT (12) accepts the alternative hypothesis* $\mathcal{H}_1$ *if, with the previous notation:*

$$\frac{\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1/2}\left(\mathbf{P}\mathbf{f}-\boldsymbol{\mu}_0\right)}{\|\boldsymbol{\theta}_1\|} > \tau^{\mathrm{lr}}.$$

In fact, the only non-linearity in the original ensemble classifier for steganalysis comes from the majority voting, which is not used in the proposed methodology. The fact that the proposed linear decision rule performed virtually identically within our experiments (see Section V-D) makes the usefulness of the non-linearity due to majority voting in the original ensemble questionable. This claim is also supported in [27] which shows that, provided the regularization parameter is correctly set, a linear classifier can be constructed with the same performance as that of the original ensemble. This indicates that the fusion of base learners in the ensemble acts as a regularization, which seems to help the detection more that the non-linear majority vote rule.

Finally, from Proposition 1, one can also compute the decision threshold $\tau^{\mathrm{P_E}}$ that minimizes the total probability of error under equal Bayesian priors, $P_E = 1/2\left(\alpha_0 + 1 - \beta(\alpha_0)\right)$ as follows, see details in Appendix 1:

Corollary 1. *The threshold given by:*

$$\tau^{\mathrm{P_E}} = \frac{\|\boldsymbol{\theta}_1\|}{2}, \tag{16}$$

minimizes the total probability of error P_E*. Using the threshold given in Equation (16), the power and the false-alarm probability of the optimal LRT, at the minimal* P_E*, are given by:*

$$\alpha_0 = \beta_{\delta^{\mathrm{lr}}} = 1 - \Phi\left(\frac{\|\boldsymbol{\theta}_1\|}{2}\right). \tag{17}$$

C. Comparison with Original Ensemble Classifier for Steganalysis

First of all, it worth noting that the proposed methodology fundamentally differs from the majority voting rule originally proposed for the FLD ensemble for the following three reasons. First, the covariance between the base learners is taken into account. Second, while the majority voting gives the same weight to all base learners, the proposed framework, as well as some other prior works [28], allows giving more importance to base learners that better distinguish the two classes. Last, as stated above, the proposed optimal LRT is a linear classifier. The main advantage of the proposed LRT is that, thanks to the statistical model detailed in Section III-A, the performance of our test is established in Proposition 1. Hence, one can not only design a test that guarantees a given false-alarm probability but also compute the associated threshold as well as the highest detection performance that can be expected. Moreover, the theoretically expected performance of the test can be compared with the empirically obtained one (see Section V-D).

Additionally, we also adjusted the search for the optimal parameters, the number of features used by each base learner d_{sub} as well as the number of base learners L, see also discussion in Section III-D. The original FLD ensemble minimizes the total probability of error under equal Bayesian priors, $P_E = 1/2\left(P_{MD} + P_{FA}\right)$, with P_{MD} and P_{FA} the probability of missed detection and false alarm, respectively (see the formal definitions in Section III-A). Since we proposed a statistical model for the ensemble classifier, we also use it to determine the optimal values of the parameters d_{sub} and L in order to maximize the theoretically established detection performance, see Proposition 1. In fact the theoretical performance only depends on the expected norm of the normalized projection vector $\|\boldsymbol{\theta}_1\|$, see Equation (10). Hence, we proposed to maximize directly $\|\boldsymbol{\theta}_1\|$ when searching for optimal values of d_{sub} and L, instead of determining these parameters by maximizing the empirical total probability P_E as proposed in the original version of ensemble classifier.

Finally, we wish to emphasize that the computational cost of the proposed LRT is very similar to the original ensemble. We note that the training of a single FLD base learner requires to evaluate the scatter matrix denoted $\widehat{\boldsymbol{\Sigma}}_c + \widehat{\boldsymbol{\Sigma}}_s$ in Equation 3 and its inversion. These two operations are respectively of complexity $\mathcal{O}(N^{\mathrm{trn}}d_{\mathrm{sub}}^2)$ and $\mathcal{O}(d_{\mathrm{sub}}^3)$. Hence, for a fixed value of the parameters d_{sub} and L the training complexity is $\mathcal{O}(N^{\mathrm{trn}}Ld_{\mathrm{sub}}^2 + Ld_{\mathrm{sub}}^3)$, where N^{trn} is the size of the training set. To evaluate the performance through the mean projection $\|\boldsymbol{\theta}_1\|$ the proposed methodology requires to estimate the covariance matrix $\boldsymbol{\Sigma_0}$ between base learners' projections and then inverting this matrix. These operations are, respectively, of complexity $\mathcal{O}(N^{\mathrm{cv}}L^2)$ and $\mathcal{O}(L^3)$, with N^{cv} the size of the cross-validation set (in our implementation, the training set is, by default, split into subsets of equal size for training and cross validation). Hence, the proposed methodology is slightly more computationally expensive. Note, however, that the difference is of order $\mathcal{O}(N^{\mathrm{cv}}L^2) + \mathcal{O}(L^3)$, which is rather negligible compared the to overall complexity of the ensemble classifier, which is $\mathcal{O}(N^{\mathrm{trn}}Ld_{\mathrm{sub}}^2 + Ld_{\mathrm{sub}}^3)$ since usually $d_{\mathrm{sub}} >> L^2$. As for the original ensemble, the overall complexity remains much smaller than others popular machine learning algorithms (*e.g.* SVM).

D. Implementation of the Proposed Methodology

The implementation of the proposed LRT is similar to the original ensemble classifier as detailed in [14]. Here, we emphasize the main implementation differences. The flowchart in Figure 1 represents the proposed implementation for the training (and testing) with the proposed methodology. First, as in [14], the features are divided into two subsets for training and testing, both equal in size by default. The testing set is used to compute the numerical results in this paper, see Figure 1, but of course, in "real life" all the samples will be used for training, the testing set will be provided by data to classify for which the "label" (either cover or steganographic images) remains unknown. Then, the training of each base

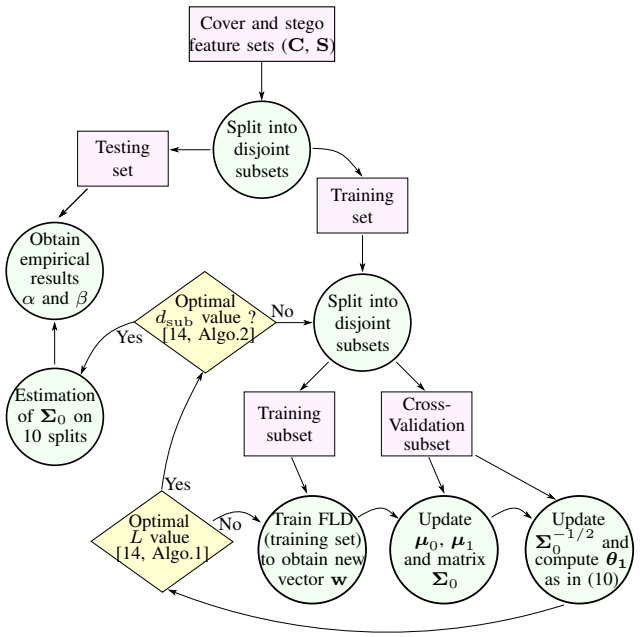


Fig. 1: Schematic representation describing the implementation of the proposed methodology. Note that rectangles represent data, circles are associated with operations, and rhombuses represent loops and associated tests.

learner is performed only on a subset of the training set, while the remaining subset of the training set is used for cross-validation to measure the performance of the base learners. In contrast to the original ensemble, we did not use bootstrapping but simply split the training set into two subsets of equal size (hence each containing one quarter of the entire feature set). All these subsets are kept disjoint in both the original ensemble as well as the proposed method.

As briefly discussed in Section III-C, we use the proposed statistical model for the ensemble classifier in order to search for the optimal values of the parameters d_{sub} and L. To this end, one only needs to measure the base learners' projection covariance $\mathbf{\Sigma_0}$ and the expectations $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ under each hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Hence, for a fixed value of d_{sub}, each base learner is added to the ensemble as follows (see the bottom row of Figure 1): (1) a randomly selected subset of features of size d_{sub} is selected (not shown in Figure 1) (2) the training set is used to obtain the projection vector $\mathbf{w}$, see Equation (3), from an FLD base learner (each base learner is trained as a usual FLD except that we do not need the threshold value b); (3) the cross-validation set is used, with the projection vector $\mathbf{w}$, to update the base learners' projection covariance $\mathbf{\Sigma}_0$ and expectations $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. Once these values have been updated, it is straightforward to re-compute $\|\boldsymbol{\theta}_1\|$ using Equation (10). Because we reused the search method used to find the parameters $d$sub and L from the original ensemble classifier [14] (shown by rhombuses in Figure 1), we refer the reader to this paper for further details.

It also worth noting that in the original ensemble a bootstraping is done so that, for each added base learner, the cross-validation is performed on a different subset of the training set and all the training samples are considered to estimate the performance using the out-of-bag (OOB) estimate of the testing error. Because this can hardly be done with the base learners' projections, we change the subset used for the cross-validation only when the number of features d_{sub} is changed and keep the split of the training set unchanged to find the number of base learners L.

Finally, note that the estimation of the parameters of the multivariate normal distribution is also difficult in this context. While the estimation of expectations is rather simple, the estimation of the covariance matrices becomes difficult when the number of base learners increases: for L base learners, $1/2L(L+1)$ covariances have to be estimated. Since the estimation of the covariance matrix of projections $\mathbf{\Sigma}_0$ plays a crucial role, once the optimal values of d_{sub} and L are found, we estimate the mean $\boldsymbol{\mu}_0$ and the covariance matrix $\mathbf{\Sigma}_0$ on ten random cross-validation splits. This is especially important for the covariance matrix estimation: an accurate estimation ensures the validity of the theoretical results for low false-alarm probability α_0 around 10^{-3}, see Section V-D.

E. Extension to Steganalysis of Unknown Payload

Until now, the most powerful LRT was designed based on the assumption that the payload R was known. In practice, however, the steganalyst usually has little or no information about the payload size. In this case, the statistical problem of detecting the presence of hidden data becomes a test between the following composite hypotheses, reusing the notation from Section III-A:

$$\begin{cases} \mathcal{H}_0: & \{\mathbf{v} \sim \mathcal{P}_{\theta_0}\}, \\ \mathcal{H}_1: & \{\mathbf{v} \sim \mathcal{P}_{\theta_R},\, R \in (0,1]\}, \end{cases} \tag{18}$$

with R the payload expressed in bits per pixel (or per non-zero AC DCT coefficient for JPEG images).

The ultimate goal of steganalysis is to design a statistical test that is Uniformly Most Powerful (UMP): a test which coincides with the Most Powerful LRT whatever the embedded payload might be. Generally speaking, such a test rarely exists except when the hypotheses have a monotone Likelihood Ratio of a scalar parameter, see [26, Theorem 3.4.1]. We adopt here the assumption that, whatever the payload used for training, when testing an image, the expectations of base learners projections are increasing with the payload. This assumption is sometimes referred to as the "shift hypothesis" and was recognized for the first time in [29]. We further assumed that the covariance matrix remains the same for any (small) payload; this is reasonable for small payloads, which is the focus of this paper. With the proposed statistical model, the problem of detecting messages of unknown length can be formally written as a choice between the following composite hypotheses:

$$\begin{cases} \mathcal{H}_0: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_L)\}, \\ \mathcal{H}_R: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(f(R)\boldsymbol{\theta}_1, \mathbf{I}_L)\}, \end{cases} \tag{19}$$

with $f : [0,1] \mapsto \mathbb{R}^+$ a monotone increasing function. Roughly speaking $f(R)$ represents the impact of the payload R, how much it "pushes" all the expectations of all base learners' projections $\widetilde{\mathbf{v}}$, see Equation (9).

Proposition 2. *(1) Assuming that the model (19) holds true, then for any $\alpha_0 \in (0,1)$ the LRT (8) with the LR defined as in (12) is UMP in the class $\mathcal{K}_{\alpha_0}$ for testing hypotheses (19) provided that the decision threshold is chosen as in Equation (14).*
(2) For any payload $R \in (0,1]$ and for any $\alpha_0 \in (0,1)$, the power function of the UMP LRT (8) is given by:

$$\begin{aligned}\beta_{\delta^{\mathrm{lr}}} = \mathbb{P}_{\mathcal{H}_1}\left(\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) > \tau^{\mathrm{lr}}\right) &= 1 - \Phi\left(\tau^{\mathrm{lr}} - f(R)\,\|\boldsymbol{\theta}_1\|\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - f(R)\,\|\boldsymbol{\theta}_1\|\right)\end{aligned} \tag{20}$$

Proof. Proposition 2 is proved in Appendix A. □

It is important here to stress that, in practice, the function $f(R)$ is not used in the proposed UMP LRT. The proposed approach requires only the knowledge of $\boldsymbol{\theta}_1$ and the covariance matrix $\boldsymbol{\Sigma}_0$, which can both be estimated for any fixed payload as in the case of simple hypotheses.
The function $f(R)$ is primarily introduced to formalize the condition under which the optimality of the proposed approach holds, which is when a given embedding scheme pushes the projections of all base learners' projections along the same direction $\boldsymbol{\theta}_1$ regardless of the payload. In practice, $f(0) = 0$ and is non-linearly increasing with the payload; it behavior depends on the steganographic embedding algorithm, its content adaptivity, the embedding domain (spatial or JPEG), and many other attributes.
Besides, since the power of the proposed UMP LRT (20) depends on $f(R)$, in the case of an unknown payload, R, the power is also unknown. This expression is provided in order to (a) contrast it with the case of simple hypothesis (15) and (b) fully characterize the statistical properties of the UMP LRT.

Note that there exist other possibilities to construct a test that is optimal regardless of the payload. It has been proposed in [5] to use the Local Asymptotic Normality [33] in order to design an asymptotically UMP test around a chosen payload $R^\star$. This approach can however hardly be used for the problem addressed in this paper because 1) we explicitly wish to address the case in which the payload is unknown and no prior information on this is available and, 2) it is hardly possible to indefinitely increase the number of base learners. A more practical approach has been proposed in [34] by training the classifier with a mixture of payloads. The author has shown that the best results are obtained using a uniform distribution of payload. This approach is contrasted with the proposed UMP test in Section V-E for two different embedding algorithms. Note that we tested many other embedding algorithms but, for brevity, only two are presented. The results obtained show the relevance of the proposed test. While the proposed UMP test performs very well, either when trained on a mixed payload or with a wrong payload, the original ensemble with majority voting performs significantly worse, see Section V-E.

IV. Optimal Criteria and Design of Statistical Tests for Multi-Class Steganalysis

The previous Section III focused on binary hypothesis testing, that is when the steganalysts' goal is to detect a specific embedding scheme. The goal of this section is to extend the proposed methodology to classification of multiple steganographic algorithms. Before presenting the proposed optimal statistical test, let us first formally state the problem of multi-class classification (or multiple hypotheses testing) and present the criterion of optimality that will be used, and that has never been studied for steganalysis.

A. Multi-class Classification: Problem Statement

In the context of multi-class classification a set of M different steganographic algorithms is suspected by the steganalyst. As in Sections III-A – III-B, let us assume that the payload R is known and that the base learners have been trained and their projection vectors normalized, see Equation (9). Hence, we have a set of $M+1$ possible hypotheses:

$$\begin{cases} \mathcal{H}_0 & : \{\widetilde{\mathbf{v}} \sim \mathcal{N}(0\,,\,\mathbf{I}_L)\} \\ \mathcal{H}_1 & : \{\widetilde{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\theta}_1\,,\,\mathbf{I}_L)\} \\ \vdots & \quad\vdots \\ \mathcal{H}_M & : \{\widetilde{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\theta}_M\,,\,\mathbf{I}_L)\}\,, \end{cases} \tag{21}$$

where, again, $\widetilde{\mathbf{v}}$ are the normalized base learners' projections defined in (9) and $\boldsymbol{\theta}_m\,,\, m = \{1,\dots,M\}$, represents the expected value of the normalized base learners' projections under hypothesis $\mathcal{H}_m$.

Remark 2. *It is worth noting that the hypotheses testing problem (21) is similar to the detection of signal in Gaussian noise, which has been extensively studied but seldom within the framework of hypothesis testing, such as in [35]–[37].*

Remark 3. *Two main settings have been used for training the ensemble: 1) gather FLD base learners when training $\mathcal{H}_0$ against $\mathcal{H}_m$, giving us M sets of base learners, 2) train an ensemble for each pair $\mathcal{H}_k$ against $\mathcal{H}_m$, with $k \neq m$, giving us $M \times (M+1)/2$ sets of FLD base learners. Since the second approach performs better, it is the only one used in this paper.*

The problem of multi-class classification is usually addressed using a vector decision function $\boldsymbol{\delta}^{\mathrm{mc}} : \mathbb{R}^L \to \{0,1\}^{M+1}$ defined by $M+1$ "sub-decision functions" $\boldsymbol{\delta}^{\mathrm{mc}} = (\delta_0^{\mathrm{mc}},\dots,\delta_M^{\mathrm{mc}})^{\mathrm{T}}$ such that hypothesis $\mathcal{H}_m\,,\, m = \{0,\dots,M\}$, is accepted if $\delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}}) = 1$ and $\forall \widetilde{\mathbf{v}} \in \mathbb{R}^L\,, \sum_{m=0}^{M} \delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}}) = 1$.
With this notation, let us formally define the following probabilities, useful for evaluating the average performance of a statistical multi-class (multiple hypotheses) test:

$$\begin{cases} \alpha_0 = P(FA) = \mathbb{E}_{\mathcal{H}_0}\left(1 - \delta_0^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right) \\ P_{\mathcal{H}_m}(MD) = \mathbb{E}_{\mathcal{H}_m}\left(\delta_0^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right) \\ \beta_m = P_{\mathcal{H}_m}(CC) = \mathbb{E}_{\mathcal{H}_m}\left(\delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right). \end{cases} \tag{22}$$

Here, $P(FA)$ corresponds to the probability of false-alarm, $P_{\mathcal{H}_m}(MD)$ denotes the probability of missed detection under hypothesis $\mathcal{H}_m$, and $P_{\mathcal{H}_m}(CC)$ represents the probability of correct classification under hypothesis $\mathcal{H}_m$ with the probability of erroneous classification defined by $\alpha_m = 1 - \beta_m = P_{\mathcal{H}_m}(EC) = \mathbb{E}_{\mathcal{H}_m}\left(1 - \delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right)$.

B. Minimax criterion and Optimal Minimax Test

The Neyman–Pearson criterion of optimality is no longer usable for multi-class classification. A natural extension of

the Neyman–Pearson approach for multi-class classification is to seek a test that, first, guarantees a prescribed false alarm probability α_0 and, second, maximizes the worst correct-classification probability β_m with respect to all possible alternative hypotheses $\mathcal{H}_1, \dots, \mathcal{H}_M$. The idea behind focusing on the worst case is similar to the example given in [38]. Consider two tests, one that provides a 60% correct classification probability for all alternative hypotheses and one that has a correct classification probability of 80% for the first half of the alternatives hypotheses and a correct classification probability of 45% for the second half of alternative hypotheses. Which test is more desirable to use in practice? It is argued in [38] that the first test is more stable. Another crucial argument is that the steganographer may choose the embedding method that the steganalyst detects the worst: one for which the correct classification probability is 45%.

In hypothesis testing theory, a test that achieves a given false alarm probability α_0 while maximizing the worst case correct-classification probability β_m, is referred to as the "constrained minimax test" [39]. Formally, for solving the detection problem defined in (21), let us define the class of all tests with a false-alarm probability bounded by α as:

$$\mathcal{D}_{\alpha_0} = \Bigg\{ \boldsymbol{\delta} : \mathbb{R}^L \to \{0,1\}^{M+1}, \sum_{m=0}^{M} \delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}}) = 1, \tag{23}$$

$$\mathbb{E}_{\mathcal{H}_0}\left(1 - \delta_0^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right) \leq \alpha \Bigg\}. \tag{24}$$

Among all the tests in (23), we try to find a test that maximizes the smallest correct-classification probability β_m, or, equivalently, minimizes the maximal α_m. Hence, the test $\boldsymbol{\delta}^{\mathrm{mc}} = (\delta_0^{\mathrm{mc}}, \dots, \delta_M^{\mathrm{mc}})^{\mathrm{T}}$ is formally said to be an optimal constrained minimax test if, for any other decision function $\boldsymbol{\delta}^{\mathrm{mc}\prime} = \left(\delta_0^{\mathrm{mc}\prime}, \dots, \delta_M^{\mathrm{mc}\prime}\right)^{\mathrm{T}}$, $\boldsymbol{\delta}^{\mathrm{mc}}$ satisfies:

$$\max_{m>0} \mathbb{E}_{\mathcal{H}_m}\Big(1 - \delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}})\Big) \leq \max_{m>0} \mathbb{E}_{\mathcal{H}_m}\left(1 - \delta_m^{\mathrm{mc}\prime}(\widetilde{\mathbf{v}})\right). \tag{25}$$

Note that while the usual minimax criterion aims at finding a test that minimizes the maximal α_m (22) (over all $M+1$ hypotheses), the constrained minimax test introduces a constraint on the false-alarm probability α_0, which is similar to the Neyman–Pearson criterion of optimality for the binary case.

A practical way to find the constrained minimax test follows from the following theorem [39].

Theorem 1. *The test $\boldsymbol{\delta}^{\mathrm{mc}} = (\delta_0^{\mathrm{mc}}, \dots, \delta_M^{\mathrm{mc}})^{\mathrm{T}}$ with the assignments*

$$\begin{cases} \delta_0^{\mathrm{mc}}(\widetilde{\mathbf{v}}) = 1 \text{ if } \max\limits_{m\in\{1,\dots,M\}} w_m + \boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}} \leq \tau^{\mathrm{mc}} \\ \delta_k^{\mathrm{mc}}(\widetilde{\mathbf{v}}) = 1 \text{ if } \max\limits_{m\in\{1,\dots,M\}} w_m + \boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}} = w_k + \boldsymbol{\theta}_k^{\mathrm{T}}\widetilde{\mathbf{v}} > \tau^{\mathrm{mc}} \end{cases} \tag{26}$$

is an optimal constrained minimax test provided the weights w_m are such that the following necessary and sufficient conditions hold:
1) $\boldsymbol{\delta}^{\mathrm{mc}} \in \mathcal{D}_{\alpha_0}$;
2) all the erroneous classification probabilities are equal:

$$\mathbb{E}_{\mathcal{H}_1}\Big(1 - \delta_1^{\mathrm{mc}}(\widetilde{\mathbf{v}})\Big) = \dots = \mathbb{E}_{\mathcal{H}_M}\Big(1 - \delta_M^{\mathrm{mc}}(\widetilde{\mathbf{v}})\Big). \tag{27}$$

Such a test is referred to as an "equalizer test" in the sense that $\forall m \in \{1, \dots, M\}$, $P_{\mathcal{H}_m}(EC) = \mathbb{E}_{\mathcal{H}_m}\left(1 - \delta_m^{\mathrm{mc}}(\widetilde{\mathbf{v}})\right) = \alpha^{\mathrm{mc}}$.

We note that the log-likelihood ratio between the hypotheses $\mathcal{H}_m$ and $\mathcal{H}_0$ can be written as $\boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}}$, see (12) and Appendix A. Hence, the term $w_m + \boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}}$ corresponds to the likelihood ratio between hypotheses $\mathcal{H}_m$ and $\mathcal{H}_0$ scaled by e^{w_m}.

Remark 4. *A formal proof of Theorem 1 can be found in [39]. One way way to understand this theorem is that if all probabilities α_m were not equal, then one could randomly accept the k-th hypothesis, with $k = \arg\max_{m\in\{1\dots M\}} \alpha_m$ (the one with the highest erroneous classification rate), instead of the j-th hypothesis, with $j = \arg\min_{m\in\{1\dots M\}} \alpha_m$ (which has the smallest erroneous classification rate) to design a test with a smaller maximal erroneous classification probability.*

Remark 5. *It also worth noting that the proposed minimax test coincides with the Bayesian test for the testing problem (21), see [25, Chap. 6] when the prior probabilities $p_0, \dots, p_M$ satisfy $w_m = \log(p_m/p_0)$. The prior distribution, which corresponds to the minimax test, is sometimes referred to as the "least favorable distribution" since it is the one for which the correct classification probability is the smallest, see [26, Chapter 3.8].*

C. Discussion and Implementation of Proposed Minimax Test

Before discussing the results of the proposed minimax test, in Section V-F, let us first recall that such a criterion of optimality, that guarantees a false-alarm rate for multi-class case, has never been studied in steganalysis.

Let us also briefly describe the implementation of the proposed minimax test which is depicted in Figure 2. Again, we have chosen to train an ensemble for each pair $\mathcal{H}_k$ against $\mathcal{H}_m$, with $k \neq m$, see Remark 3 and the training step in Figure 2. Then all the base learners are aggregated: that is the features to be classified are projected onto projection vectors of all base learners from all classifiers; this in fact does not cause computation issue even though the ensembles may have different base learners which may be build on subset of features of different dimension.

Again, as in the binary testing case discussed in Section III-D, we note that in fact, the proposed methodology only needs to estimate the base learners' projection expectations and covariances. Hence, from the aggregated projection vectors, it is straightforward to estimate the covariance matrix $\boldsymbol{\Sigma}_0$ and the expectation $\boldsymbol{\mu}_0$ on the cross-validation subset. This ensures that all projections are normalized and that the proposed model (21) holds true. This is the normalization step in Figure 2. Finally, the means of the normalized projections are computed from the cross-validation subset for each alternative hypothesis. From this last step, the application of the proposed minimax test is straightforward. Note that because of the higher number of projections involved in this case, the accuracy of the covariance matrix estimate is extremely important, hence the necessity of estimating it with many different cross-validation splits described in Section III-C.

Note that the classifiers trained without $\mathcal{H}_0$ are used in

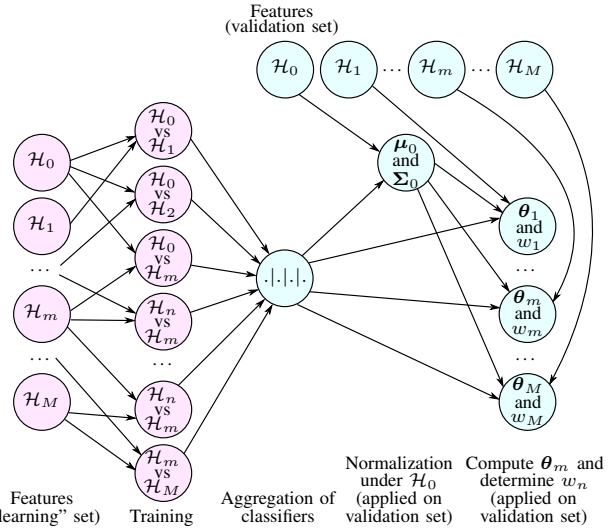


Fig. 2: Schematic representation describing the application of proposed multi-hypotheses minimax test.

the same way, and, from experiments, we noted that their use substantially helps to decrease the misclassification of embedding schemes (or alternatives hypotheses).

It should be noted that the weights w_m, which are required to "equalize the erroneous classification probability" (27), are computed numerically because it is hardly possible to find an analytic formula for the weights w_m that "equalize" erroneous classification probabilities. Indeed, let us define:

$$\boldsymbol{\Lambda}^{\mathrm{mc}} = \left(\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}}, \ldots, \boldsymbol{\theta}_M^{\mathrm{T}}\widetilde{\mathbf{v}}\right)^{\mathrm{T}},$$

which correspond to the M values of the log-likelihood ratios between hypotheses $\mathcal{H}_m$, $m \in \{1, \ldots, M\}$, and $\mathcal{H}_0$. It is straightforward to establish from the statistical model of normalized base learners' projections $\widetilde{\mathbf{v}}$ that $\boldsymbol{\Lambda}^{\mathrm{mc}}$ follows, under hypothesis $\mathcal{H}_m$, a multivariate normal distribution with expectation $\mathbb{E}_{\mathcal{H}_m}\left(\boldsymbol{\theta}_n^{\mathrm{T}}\widetilde{\mathbf{v}}\right) = \boldsymbol{\theta}_n^{\mathrm{T}}\boldsymbol{\theta}_m$ and covariance $Cov\left(\boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}}, \boldsymbol{\theta}_n^{\mathrm{T}}\widetilde{\mathbf{v}}\right) = \boldsymbol{\theta}_m^{\mathrm{T}}\boldsymbol{\theta}_n$. Even though some accurate approximations of multivariate normal probabilities may be found [40], [41], analytically computing the threshold that guarantees that $\mathbb{P}_{\mathcal{H}_0}\left(\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}} \leq \tau^{\mathrm{mc}} \cap \ldots \cap \boldsymbol{\theta}_M^{\mathrm{T}}\widetilde{\mathbf{v}} \leq \tau^{\mathrm{mc}}\right) = 1 - \alpha_0$ is hardly possible. Similarly, the correct classification probability $\mathbb{P}_{\mathcal{H}_m}\left(\bigcap_{n\neq m} \boldsymbol{\theta}_n^{\mathrm{T}}\widetilde{\mathbf{v}} < \boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}} \cap \boldsymbol{\theta}_m^{\mathrm{T}}\widetilde{\mathbf{v}} > \tau^{\mathrm{mc}}\right)$ requires integrating a multivariate normal pdf over a complicated domain, which is quite difficult to achieve even numerically.

We thus opted to compute the weights numerically using a gradient-free Nelder–Mead (downhill simplex) method [42] on the training samples. The initial weights are set to zero while the threshold is fixed at $\tau^{\mathrm{mc}} = \Phi^{-1}(1 - \alpha_0/M)$, which is easily shown to guarantee the prescribed false-alarm probability α_0 in the case of orthogonal vectors $\boldsymbol{\theta}_m$. Then, on the cross-validation subset, we numerically compute the "weights" w_m that (1) guarantee the prescribed false-alarm rate α_0 and (2) equalize the erroneous-classification probability α_m, $\forall m \in \{1, \ldots, M\}$. This was done using *Matlab* built-in optimization functions by minimizing a weighted sum of the two objective functions.

V. Numerical Simulations and Results

A. Common core of all experiments

All results presented in this paper are obtained on BOSSbase 1.01 [22] containing $10,000$ 512×512 gray-scale images. The detection errors are always computed by averaging over 10 different random database splits. Four spatial domain embedding schemes were used: HUGO [31] with bounding distortion (HUGO-BD) implemented using the Gibbs construction [32], the Wavelet Obtained Weights (WOW) [30] algorithm, the spatial version of UNIversal WAvelet Relative Distortion (S-UNIWARD) [1], and the recent scheme based on statistical detectability [16], [17]. For spatial domain steganalysis, we used the second-order Subtractive Pixel Adjacency Matrix (SPAM) [12] feature set of dimensionality 686, the Spatial Rich Model (SRM) [43] with dimensionality $34,671$ and its selection-channel-aware version (maxSRMd2) [44].

Several embedding methods for JPEG domain have also been included in our tests, namely, nsF5 [45], the Entropy-Based Steganography (EBS) [47], the Uniform Embedding Distortion (UED) [46], and the JPEG domain version of UNIWARD referred to as J-UNIWARD [1]; the two last being the state-of-the-art in JPEG domain steganography that does not use side information. For JPEG domain embedding scheme with side information we used the Perturbed Quantization (PQ) [45], the side-informed version of EBS (SI-EBS) [47], and the side-informed version of UNIWARD, SI-UNIWARD [1]. The four different feature sets that have been used for JPEG image steganalysis are the Cartesian-calibrated JPEG Rich Model (CC-JRM) [48] with $22,510$ features, the $\mathcal{CF}^{\star}$ [14] feature set with $7,850$ features, the spatial rich model with fixed quantization (SRMQ1) [43] of dimensionality $12,753$, and the union of SRMQ1 and CC-JRM, referred to as JSRM [48], whose dimensionality is $35,263$. All feature extractors used in this paper and most embedding algorithms can be downloaded from the DDE website at http://dde.binghamton.edu/download.

Note that the payload is measured in bits per pixel (bpp) for spatial domain steganography and in bits per non-zero AC coefficients (bpnzAC) for JPEG domain steganography. The JPEG images were created using the imwrite function from *Matlab* with quality factors 75 and 90 for both side-informed and non-side informed schemes. The precover was either the uncompressed image (for SI-UNIWARD and SI-EBS) or a JPEG image compressed with quality factor 90 and 100 (for the PQ algorithm). Finally, note that for HUGO-BD, the switch T was set to 255 to remove the weakness identified during the BOSS contest. The stabilizing constant σ for UNIWARD was set to 1 for the spatial version and to 10^{-6} for JPEG versions to prevent the attack proposed in [1].

Finally, note that when the detection accuracy is measured as the total probability of error under equal Bayesian priors, $P_E = 1/2\,(P_{MD} + P_{FA})$, we used the threshold $\tau^{\mathrm{P_E}}$ as given in Corollary 1, Equation (16).

B. Experiment and Comparison on Simulated Data

To verify the sharpness of the theoretical results, we selected the covariance matrix corresponding to five randomly selected base learners for the experiment on multiple hypothesis testing

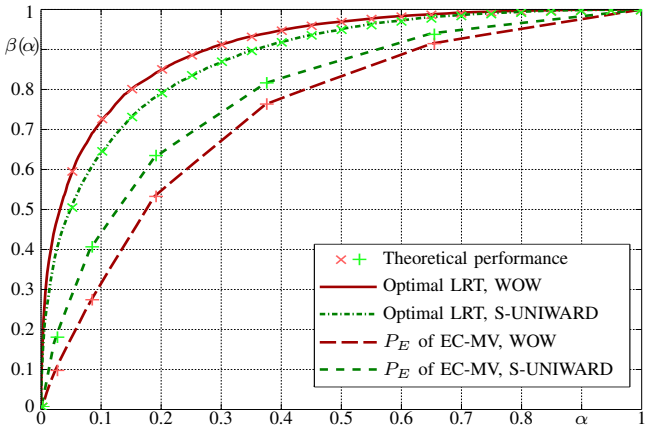


Fig. 3: ROC curves for the ensemble classifier implemented with majority voting and the proposed optimal LRT comparing empirical and theoretical results: toy example with five randomly chosen base learners (see the text for more details).

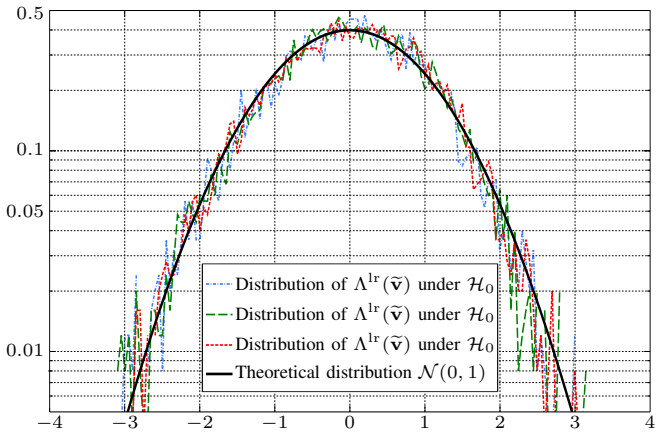


Fig. 4: Comparison between the theoretical normal distribution and the empirical distribution of the proposed LR under $\mathcal{H}_0$ on one half of BOSSbase [22] used for testing.

reported in Table X. It used the SRM features and three different embedding schemes, WOW, S-UNIWARD, and HUGO-BD, all with payload $R = 0.4$. The estimated covariance under hypothesis $\mathcal{H}_0$ is given by:

$$\mathbf{\Sigma_0} = \begin{pmatrix} 0.9823 & 0.5524 & 0.5218 & 0.1734 & 0.2724 \\ 0.5524 & 1.1416 & 0.6179 & 0.3140 & 0.4191 \\ 0.5218 & 0.6179 & 1.2359 & 0.2996 & 0.6512 \\ 0.1734 & 0.3140 & 0.2996 & 0.8243 & 0.4635 \\ 0.2724 & 0.4191 & 0.6512 & 0.4635 & 1.4514 \end{pmatrix},$$

which clearly shows a very high correlation between the base learners' projections.

Then, a total of 10^7 samples following a multivariate normal distribution with this covariance have been randomly generated for each hypothesis and the proposed optimal binary test has been applied and compared with the theoretically established power function $\beta_{\delta^{\mathrm{lr}}}$, see (15). Figure 3 shows the ROC curves for the binary test of $\mathcal{H}_0$ (covers) versus hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ (corresponding to WOW and S-UNIWARD at $R = 0.4$). These results indicate that the proposed optimal LRT performs almost exactly as theoretically established.

For comparison, we also include the results obtained with the original ensemble classifier with the majority voting (denoted EC-MV). Note that for a very low L, the optimal LRT performs much better, see also Figure 6. Most importantly, however, establishing the statistical properties of the EC-MV detector is quite expensive. Even in this toy example, drawing the ROCs required 32 evaluations of the multivariate normal cumulative distribution function (cdf) to establish the probability of having at least $N_c \in 0, \ldots, 5$ base learner votes. More precisely, in this toy example, one needs to compute the probability that exactly N_c base learners among 5 classify an image as stego. Since there exists $\binom{5}{N_c}$ possibilities that exactly N_c base learners classify an image as stego, this requires evaluating the same number of time a normal mutivariate cdfs. For establishing the performance of the original ensemble classifier with a variable threshold on the number of votes, this should be done for $N_c \in 0, \ldots, 5$.

This example can be extended easily to L base learners when one wants to establish the probability of false alarm and the power as a function of the number of base learner votes using a multivariate Gaussian model or, in fact, any statistical model. For each threshold N_c on the number of base learners that classify an image as stego, there is $\binom{L}{N_c}$ possibilities. So if one wants to establish the power and the false-alarm probability for all possible values of N_c, so that it can select the one that suits it needs in terms of the false-alarm probability for instance, this requires a total of 2^L possibilities.

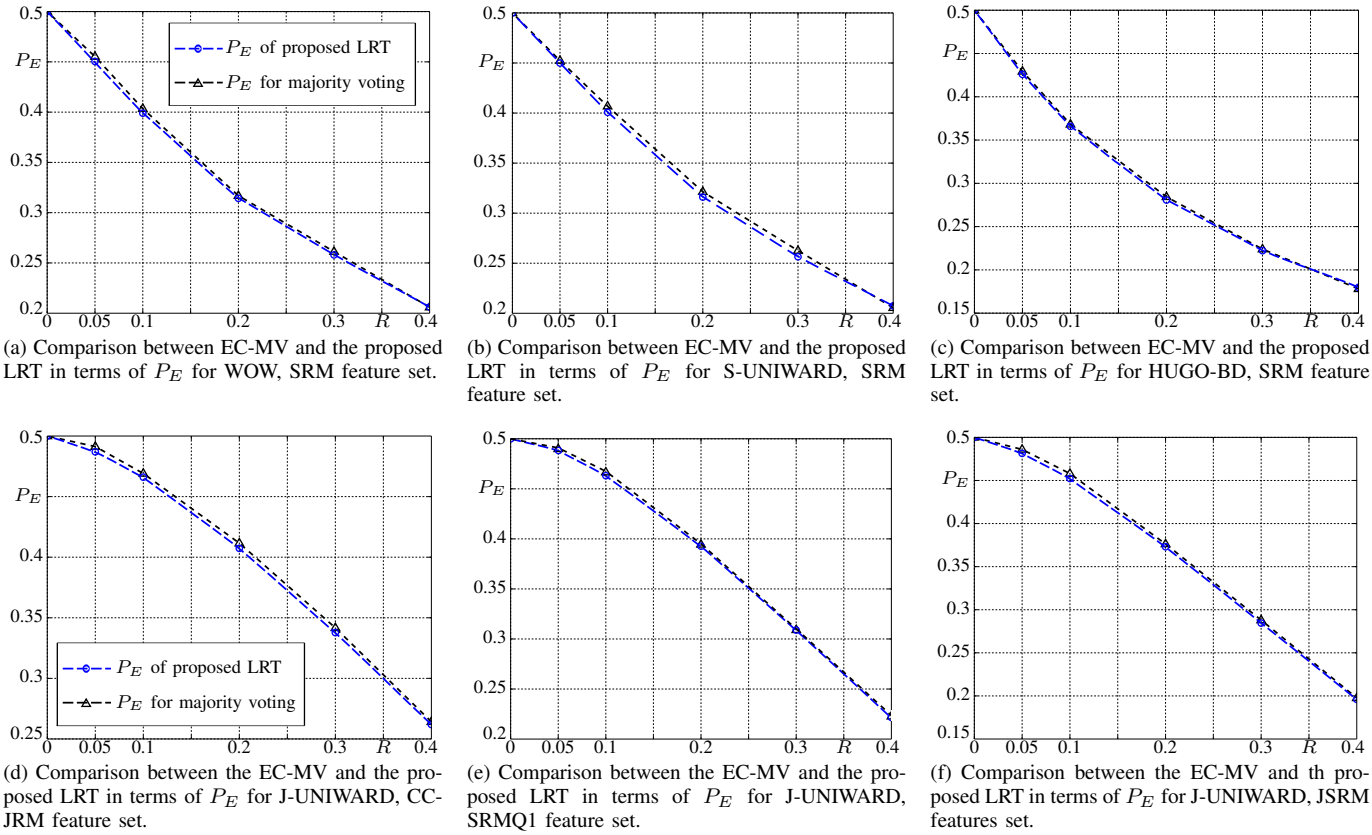
A similar complication occurs when optimizing the threshold of each base learner to maximize the detection accuracy. This will also be extremely time consuming to compute using a statistical model of base learners for the same reason. A numerical joint optimization over L thresholds of all base learners will certainly be computation expensive as well.

C. Relevance of the Proposed Model

A wide range of numerical experiments have been performed to confirm the assumption of mutivariate normality of base learners' projections. For brevity and clarity, in this section we include numerical results that support the claim that the proposed LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$ follows a normal distribution with 0 mean and unit variance under $\mathcal{H}_0$, which is crucial to guarantee a prescribed false-alarm probability, see Proposition 1.

First, Figure 4 shows a comparison between the theoretical Gaussian distribution of the LR, $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$, and the empirical distribution obtained with optimal d_{sub} and L for three different algorithms (WOW [30], S-UNIWARD [1], and HUGO-BD [31]). This provides visually interpretable results that show the match between the empirical distribution and the theoretical model.

Next, Tables I ans II show the χ^2 goodness-of-fit (GOF) score obtained when comparing the empirical LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$ distribution and one with the assumed pdf $\mathcal{N}(0,1)$ under hypothesis $\mathcal{H}_0$. Because of the rather high number of testing data, we have used 100 bins to apply the χ^2 GOF test. Hence, this score should follow a χ^2 distribution with 100 degree of freedom, that has an expectation of 100 with variance 200, with higher scores indicating a larger deviation from the assumed model.



(a) Comparison between EC-MV and the proposed LRT in terms of P_E for WOW, SRM feature set.

(b) Comparison between EC-MV and the proposed LRT in terms of P_E for S-UNIWARD, SRM feature set.

(c) Comparison between EC-MV and the proposed LRT in terms of P_E for HUGO-BD, SRM feature set.

(d) Comparison between the EC-MV and the proposed LRT in terms of P_E for J-UNIWARD, CC-JRM feature set.

(e) Comparison between the EC-MV and the proposed LRT in terms of P_E for J-UNIWARD, SRMQ1 feature set.

(f) Comparison between the EC-MV and th proposed LRT in terms of P_E for J-UNIWARD, JSRM features set.

Fig. 5: Comparison between the proposed LRT and the majority vote decision rule for different spatial domain embedding schemes, top row, and for J-UNIWARD with different feature sets, bottom row.

Feature	$D = 8000$				$D = 64000$			
$d_{\text{sub}} =$	400	800	1600	3200	400	800	1600	3200
$L = 50$	413.0	355.6	347.2	645.7	410.0	339.1	307.6	449.3
$L = 100$	354.8	326.7	269.8	639.4	313.1	312.1	260.3	396.8
$L = 200$	287.0	243.8	269.2	562.0	234.7	262.7	227.7	334.0
$L = 400$	213.7	217.3	231.3	485.7	208.5	211.7	204.2	307.1

TABLE I: Values of χ^2 GOF for several values of d_{sub} and L. Results obtained with MiPOD [16] and gathering SRM and maxSRMd2 [44] features.

We hypothesize that the normal distribution of base learners' projection is a consequence of CLT. Hence, the model should be more accurate when increasing the number of features D, the number of features used by each base learner d_{sub}, and the number of base learners L. Table I shows the χ^2 GOF score for the content-adaptive scheme [16] when using both SRM and maxSRMd2 features [44], to be able to increase the number of features. As shown in Table I, when the number of base learners L increases, the assumed model becomes more accurate. Similarly, when d_{sub} increases the goodness also generally increases, however, one can note that when it becomes too large the empirical distribution starts diverging significantly from the expected model. This can be explained by the fact that in this case the base learners become strongly correlated, which destabilizes the inversion of the covariance matrix $\mathbf{\Sigma}_0$.

Next, we wanted to simulate the use of a larger database, so that we would be able to learn more accurately the diversity of images in BOSSbase. To this end, Table II shows the χ^2 GOF score when the training size is increased. The results presented in Table II have been obtained using UED [46] and JSRM [48] features with the ratio of data used for training ranging from 0.3 to 0.9. These results confirm that when the diversity of images can be learned more accurately, the proposed model also becomes more accurate.

Training ratio	0.3	0.4	0.5	0.6	0.7	0.8	0.9
χ^2-GOF score	270.8	209.9	202.3	161.9	150.9	138.0	123.3

TABLE II: Values of χ^2 GOF as a function of the training ratio for UED [46] and JSRM [48] features.

Tables I and II also show the limit of the proposed model because a χ^2 GOF score of about 200 shows a non-negligible deviation from the assumed model. The proposed model is in fact not perfect and especially we note that, because the learning is carried out over a range of images, some images act as outliers because they have singular properties (very smooth content, underexposed scenes, etc. ...). We also note, see Table II that the accuracy of the model is estimated on the testing set, while parameters are estimated using the training set and that those may be slightly different.

D. Numerical Results for the Binary Case

While the main goal of the present paper is to analytically establish the statistical properties of the ensemble classifier within the proposed framework of hypothesis testing and to

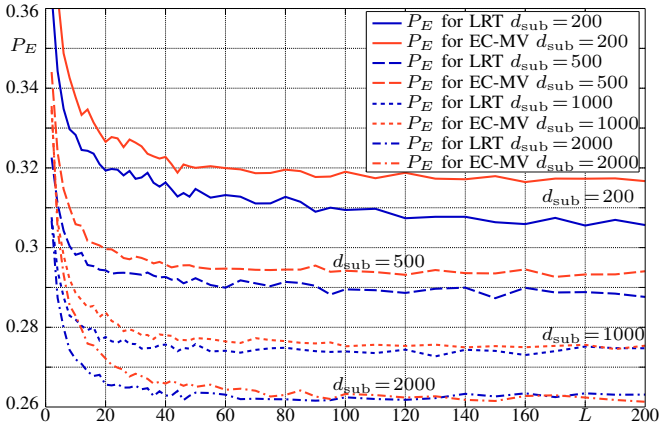


Fig. 6: Comparison between the performance of the proposed optimal LRT and the EC-MV detector as a function of L for selected values of $d_{\rm sub}$.

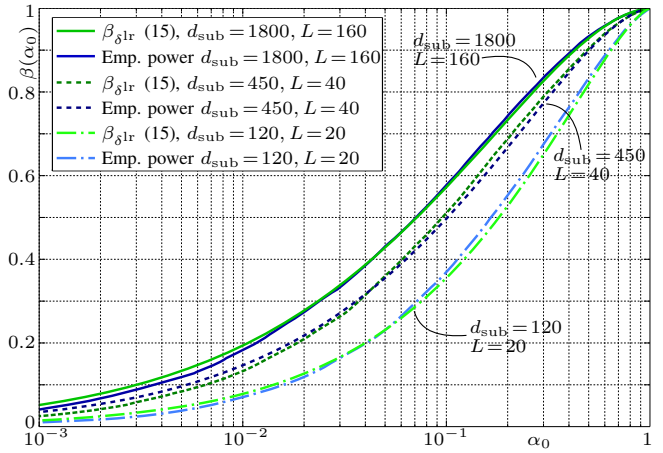


Fig. 7: Comparison between the theoretically established and empirical power function β_δ for selected values of $d_{\rm sub}$ and L.

extend its scope, it is also very important to ensure that the proposed optimal test does not perform significantly worse than the original EC-MV detector [14]. To this end, Figure 5 shows a comparison between the EC-MV and the proposed optimal LRT. These results were obtained by searching for the optimal values of the parameters $d_{\rm sub}$ and L for each detector separately, that is by minimizing P_E for the EC-MV and maximizing the theoretical power function $\beta_{\delta^{\rm lr}}$ for the proposed LRT, see Section III-C. For diversity, the top row of Figure 5 shows a comparison between the EC-MV and the proposed optimal LRT for spatial domain steganography: the same feature set (SRM) is used with three different embedding schemes: WOW in Figure 5a, S-UNIWARD in Figure 5b, and HUGO-BD in Figure 5c. The bottom row of Figure 5 shows a comparison between the proposed optimal LRT and the EC-MV for J-UNIWARD using three different feature sets: the CC-JRM in Figure 5d, the $\mathcal{CF}^\star$ in Figure 5e, and the JSRM in Figure 5f. Note that the detection performance is measured in both figures using the usual mean probability of error P_E even though the proposed LR is designed to maximize the detection power under a false-alarm probability constraint.

Figure 5 shows that the proposed optimal LRT achieves basically the same performance as the EC-MV detector. However, both detectors behave differently with respect to the parameters $d_{\rm sub}$ and L. This is demonstrated in Figure 6, which shows the total probability of error P_E as a function of L for a few fixed values of $d_{\rm sub}$. As discussed at the beginning of Section V, the proposed optimal LRT performs much better for small values of L or for small values of $d_{\rm sub}$. For large values of L and $d_{\rm sub}$ the performance of both detectors becomes very similar. The results presented in Figure 6 were obtained with the CC-JRM feature set and J-UNIWARD at payload $R = 0.4$ bpnzAC. Similar trends have been observed for other feature sets and embedding methods.

Finally, we wanted to verify the accuracy of theoretically established results, false alarm probability and power function, see Proposition 1. To this end, Figure 7 presents a comparison between the theoretically established optimal LRT power $\beta_{\delta^{\rm lr}}$, see (15), as a function of the false-alarm (ROC curve), and the empirical power function obtained on the testing set. The results were obtained using the JSRM feature set for J-UNIWARD with payload $R = 0.4$ bpnzAC. Similar results have been obtained with other feature sets and embedding methods in both the spatial and JPEG domain.

Figure 7 shows that the theoretically established power function almost perfectly matches the empirical power function. However, as discussed above, using a very high number of features for each base leaner $d_{\rm sub}$, typically beyond the optimally found parameter, would certainly decrease the accuracy of the theoretical results.

Finally, we would like to emphasize the possibility of the proposed approach to guarantee a prescribed false-alarm probability. To this end, Table III compares the empirical and the theoretical false-alarm probabilities for three different threshold values that correspond respectively to $\alpha_0 = \{2, 1, 0.5\} \cdot 10^{-3}$. The results presented in Table III have been obtained with the parameters L and $d_{\rm sub}$ always set to their optimal values.

We note that, generally, the empirical false-alarm probabilities are close the theoretical ones. However, these results have been obtained with 10 equally sized random splits of BOSSbase for training and testing, giving us a total of $50,000$ LR values. Thus, it is important to note that the very low prescribed false-alarm probability corresponds to the limit of what can be obtained within this setup as measuring a false-alarm probability of 10^{-3} corresponds to 50 images. In practice, it would be interesting to study even much lower false-alarm probabilities, but in such a case the use of the CLT is disputable as it is not relevant to model the tails of distribution.

Similarly to Table III, Figure 8 shows a comparison between the theoretically established false-alarm probability as a function of the decision threshold, $1 - \Phi(\tau^{\rm lr}) = \alpha_0$, see (14), and the empirically measured false-alarm probability obtained on the testing set. Again, for brevity, only the results obtained from J-UNIWARD with payload $R = 0.4$ bpnzAC and steganalysis using the JSRM feature set are shown. From this figure, the empirical and theoretical false-alarm probabilities are very close and similar trends can be found for other embedding methods and feature sets, see Table III.

The results presented in Figure 8 and in Table III clearly demonstrate that it is feasible in practice to accurately guarantee even a low false-alarm rate (typically around $\alpha_0 = 10^{-3}$).

Feature set	Embedding algorithm	$\alpha_0 =$ 0.2%	$\alpha_0 =$ 0.1%	$\alpha_0 =$ 0.05%
SRMQ1 [43]	WOW [30]	0.204%	0.134%	0.094%
	S-UNIWARD [1]	0.206%	0.120%	0.062%
SRM [43]	WOW [30]	0.224%	0.146%	0.100%
	S-UNIWARD [1]	0.234%	0.150%	0.086%
maxSRMd2 [44]	MiPOD [16]	0.156%	0.120%	0.094%
$\mathcal{CF}^\star$ [14]	EBS [47]	0.194%	0.112%	0.076%
	SI-EBS [47]	0.238%	0.138%	0.086%
	J-UNIWARD [1]	0.244%	0.142%	0.102%
	SI-UNIWARD [1]	0.200%	0.118%	0.078%
	nsF5 [45]	0.140%	0.100%	0.074%
	UED [46]	0.210%	0.120%	0.076%
JSRM [48]	EBS [47]	0.252%	0.172%	0.120%
	SI-EBS [47]	0.216%	0.134%	0.092%
	J-UNIWARD [1]	0.244%	0.136%	0.076%
	SI-UNIWARD [1]	0.180%	0.122%	0.084%
	nsF5 [45]	0.222%	0.142%	0.094%
	UED [46]	0.234%	0.130%	0.062%

TABLE III: Comparison between the empirical and the theoretical false-alarm probabilities for a wide-range of steganographic algorithms and feature sets.

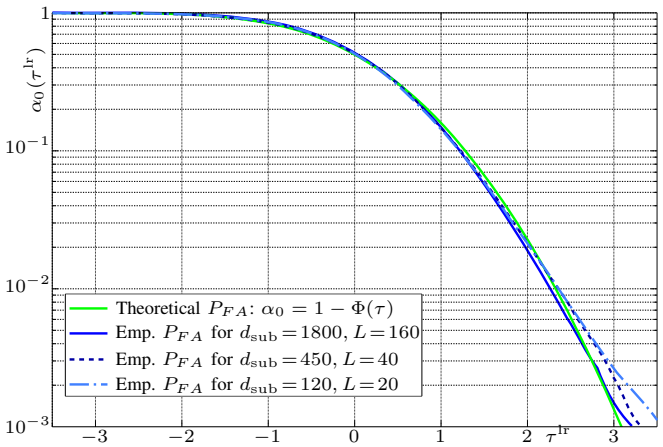


Fig. 8: Comparison between the theoretically established and the empirical probability of false-alarm as a function of the decision threshold τ.

E. Results for Composite Hypotheses: Unknown Payload

Tables IV and VI show the detection power β, or correct detection probability, Equation (7), obtained with the optimal LRT that knows the embedding rate (the clairvoyant test), the LRT when the ensemble classifiers were trained on a uniform mixture of payloads, and the proposed UMP test when the ensemble classifiers were trained with a fixed payload $R = 0.4$.

This should be contrasted with Tables V and VII also showing the detection power under the same setting with the original ensemble classifier trained to minimize P_E. While the detection power for clairvoyant detectors produces almost the same results, when either trained on payload mixture or with a fixed payload $R = 0.4$, the original ensemble performs significantly worse. This can be explained in part by the non-linearity of the majority vote, which tends to classify steganographic images with a smaller payload than the one used for training as covers. Note that here, the false-alarm probability α_0 is again set at 0.075 i.e. 7.5%. Note also that, for the original ensemble classifier, we replaced the majority voting rule so that the hypothesis $\mathcal{H}_1$ is accepted when reaching a given "number of votes" N_c determined to satisfy the false-alarm constraint on the cross-validation subset. Similar results have been obtained with the original majority voting rule and the usual minimal P_E measure of detection accuracy with other embedding schemes and feature sets.

Payload	Clairvoyant test	Payload mixture	Trained for $R = 0.4$
0.4	0.9015	0.8950	0.9007
0.3	0.6444	0.6286	0.6375
0.2	0.3399	0.3185	0.3122
0.1	0.1510	0.1473	0.1418
0.05	0.1041	0.0998	0.0984

TABLE IV: Power β, see Equation 7, of the proposed optimal UMP test for false-alarm probability $\alpha_0 = 0.075$ and UED [46] embedding scheme with $\mathcal{CF}^\star$ [14] features; d_{sub} and L set at the optimal values.

Payload	Clairvoyant test	Payload mixture	Trained for $R = 0.4$
0.4	0.9009	0.8521	0.9008
0.3	0.6403	0.5932	0.6211
0.2	0.3355	0.2495	0.2449
0.1	0.1514	0.1427	0.1113
0.05	0.1029	0.0990	0.0867

TABLE V: Power β of the original ensemble classifier with the same settings as in Table IV.

Payload	Clairvoyant test	Payload mixture	Trained for $R = 0.4$
0.4	0.5493	0.5663	0.5484
0.3	0.4538	0.4555	0.4472
0.2	0.3415	0.3340	0.3169
0.1	0.2268	0.2045	0.1866
0.05	0.1603	0.1489	0.1383

TABLE VI: Power β, see Equation 7, of the proposed optimal UMP test for false-alarm probability $\alpha_0 = 0.075$ and the content-adaptive scheme [16] with maxSRMd2 features [44]; d_{sub} and L set at the optimal values.

Payload	Clairvoyant test	Payload mixture	Trained for $R = 0.4$
0.4	0.5382	0.5641	0.5438
0.3	0.4356	0.4416	0.3902
0.2	0.3187	0.2934	0.1843
0.1	0.2064	0.1319	0.1164
0.05	0.1418	0.0896	0.0891

TABLE VII: Power β of the original ensemble classifier with same settings as in Table VI.

The results given in Tables VI and VII are also interesting because the steganalysis features utilize the knowledge of the selection channel, the probabilities of changing each pixel during embedding [44]. For such features, assuming two different payloads for training and testing may thus create an additional mismatch for the detector because the change probabilities used for feature extraction will also be different. We note that in this case the proposed LRT trained for $R = 0.4$ still achieves good performance as compared to the clairvoyant LRT, though the loss of detection accuracy is larger.

F. Extension to Multi-Class Steganalysis

The performance of the proposed constrained minimax test is shown on selected cases in Tables VIII–XI for a prescribed false-alarm rate of $\alpha_0 = 0.075$. The training was carried out as described in Section IV-C. The parameters d_{sub} and L were set to their optimal values as described in Section III-C. Tables VIII and IX present the detection performance for JPEG domain steganographic schemes using

True/accept	$\mathcal{H}_0$(Cover)	$\mathcal{H}_1$(J-UNIW)	$\mathcal{H}_2$(UED)	$\mathcal{H}_3$(EBS)
$\mathcal{H}_0$	0.9203	0.0665	0.0067	0.0065
$\mathcal{H}_1$	0.5154	0.4509	0.0170	0.0167
$\mathcal{H}_2$	0.1161	0.3835	0.4679	0.0325
$\mathcal{H}_3$	0.0465	0.4600	0.0201	0.4734

TABLE VIII: Classification accuracy for the proposed multi-class minimax test, JPEG embedding schemes, payload $R = 0.4$, JSRM features.

True/accept	$\mathcal{H}_0$(Cover)	$\mathcal{H}_1$(nsF5)	$\mathcal{H}_2$(UED)	$\mathcal{H}_3$(EBS)
$\mathcal{H}_0$	0.9202	0.0198	0.0409	0.0190
$\mathcal{H}_1$	0.0661	0.8259	0.0987	0.0093
$\mathcal{H}_2$	0.0969	0.0127	0.8228	0.0676
$\mathcal{H}_3$	0.0379	0.0063	0.1317	0.8241

TABLE IX: Classification accuracy for the proposed multi-class minimax test, JPEG embedding schemes, payload $R = 0.2$ for nsF5 and $R = 0.4$ for UED and EBS, JSRM features.

True/accept	$\mathcal{H}_0$(Cover)	$\mathcal{H}_1$(WOW)	$\mathcal{H}_2$(S-UNIW)	$\mathcal{H}_3$(HUGO-BD)
$\mathcal{H}_0$	0.9228	0.0363	0.0277	0.0131
$\mathcal{H}_1$	0.5347	0.3601	0.0879	0.0174
$\mathcal{H}_2$	0.4653	0.1524	0.3531	0.0292
$\mathcal{H}_3$	0.5022	0.0613	0.0566	0.3800

TABLE X: Classification accuracy for the proposed multi-class minimax test, spatial domain embedding schemes, payload $R = 0.4$, SRM features.

True/accept	$\mathcal{H}_0$(Cover)	$\mathcal{H}_1$(WOW)	$\mathcal{H}_2$(S-UNIW)	$\mathcal{H}_3$(HUGO-BD)
$\mathcal{H}_0$	0.9239	0.0313	0.0249	0.0199
$\mathcal{H}_1$	0.7164	0.1763	0.0828	0.0245
$\mathcal{H}_2$	0.6857	0.0841	0.2003	0.0300
$\mathcal{H}_3$	0.6964	0.0445	0.0496	0.2096

TABLE XI: Classification accuracy for the proposed multi-class minimax test, spatial domain embedding schemes, payload $R = 0.2$, SRM features.

JSRM [48] features. The results presented in Table VIII show the detection accuracy for distinguishing J-UNIWARD [1], UED [46], and EBS [47], all with payload $R = 0.4$ bpnzAC. The results presented in Table IX are the same except for J-UNIWARD replaced by nsF5 [45] with payload $R = 0.2$ bpnzAC. Interestingly, the correct classification probability for UED and EBS increased in this case. In fact, since the detection of J-UNIWARD is much harder than the detection of nsF5, the proposed equalizer test focuses mainly (in the first set of results) on maximizing the correct classification of J-UNIWARD to the detriment of UED and EBS. The probability of mis-classifying UED and EBS as J-UNIWARD is much larger in Table VIII than the probability of mis-classifying EBS and nsF5 as UED in Table IX.

The results presented in Tables X–XI show the detection accuracy for spatial domain steganography: classification of WOW [30], S-UNIWARD [1], and HUGO with bounding distortion [31], [32] using the SRM [43]. The results shown in Table X indicate that at payload $R = 0.4$ bpp the above state-of-the-art algorithms can be correctly classified with a probability as high as 35% for a false-alarm probability as low as 0.075 despite the fact that all these algorithms are very similar, in the sense that they all place the embedding changes adaptively based on content complexity, and hence are hard to discern from each other.

We note that, however, the proposed minimax test implies a higher missed-detection rate for each individual embedding scheme compared with the individual binary tests with known payload. This is natural because the test must guarantee a prescribed false-alarm rate with respect to several alternative hypotheses. For the same reason, the correct classification rate of the proposed minimax test decreased for each alternative hypothesis. This is because in the multi-class case one can mis-classify a steganographic algorithm as another one, which cannot happen in the binary case.

Finally and most importantly, Tables VIII–XI testify that the prescribed false-alarm probability of 0.075 can indeed be achieved on the testing set, which was one of the main goals of the proposed optimal minimax test.

VI. Conclusion

This paper proposes a statistical model of base learners' projections in an ensemble of linear classifiers for steganalysis of digital images. The main assumptions adopted here are that the base learners' projections follow a multivariate normal distribution and that the covariance matrix remains constant under information hiding, at least for small payloads. This statistical model is used within the framework of hypothesis testing theory to achieve the following three main goals. First, the statistical properties of the optimal LRT designed for binary case, *i.e.*, targeted steganalysis, are analytically established. Second, the LRT is extended to an optimal detector of steganography when the payload is unknown. Last but not least, the proposed framework permits extending the ensemble to multi-class steganalysis. The validity of the proposed statistical model based on which the sharpness of the theoretically established results holds has been confirmed by extensive and diverse experiments. Finally, because the proposed LRT is linear, this work also questions the usefulness of the non-linear majority vote by showing that a linear classifier can achieve roughly same performance for steganalysis using current state-of-the-art high dimensional rich feature spaces.

Acknowledgements

The authors would like to thank Vojtěch Holub for providing the code for steganographic algorithms as well as Tomáš Denemark and Vahid Sedighi for fruitful discussions.
An open-source Matlab demo code, used for this paper, is available online within download section of DDE website at dde.binghamton.edu/download/ and on the UTT-LM2S website.

References

[1] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–13, 2014.

[2] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný, "Moving steganography and steganalysis from the laboratory into the real world," in *ACM Information hiding and multimedia security, IH&MMSec'13*, 2013, pp. 45–58.

[3] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath, "Detection of hiding in the least significant bit," *Signal Processing, IEEE Transactions on*, vol. 52, no. 10, pp. 3046 – 3058, oct. 2004.

[4] R. Cogranne, C. Zitzmann, L. Fillatre, I. Nikiforov, F. Retraint, and P. Cornu, "A cover image model for reliable steganalysis," in *Information Hiding*, ser. LNCS vol.6958, 2011, pp. 178 – 192.

[5] C. Zitzmann, R. Cogranne, F. Retraint, I. Nikiforov, L. Fillatre, and P. Cornu, "Statistical decision methods in hidden information detection," in *Information Hiding*, ser. LNCS vol.6958, 2011, pp. 163 – 177.
[6] R. Cogranne, C. Zitzmann, F. Retraint, I. V. Nikiforov, P. Cornu, and L. Fillatre, "A local adaptive model of natural images for almost optimal detection of hidden data," *Signal Processing*, vol. 100, pp. 169 – 185, July 2014.
[7] T. H. Thai, F. Retraint, and R. Cogranne, "Statistical detection of data hidden in least significant bits of clipped images," *Signal Processing*, vol. 98, pp. 263 – 274, May 2014.
[8] R. Cogranne and F. Retraint, "An asymptotically uniformly most powerful test for LSB matching detection," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 3, pp. 464–476, March 2013.
[9] C. Zitzmann, R. Cogranne, L. Fillatre, I. Nikiforov, F. Retraint, and P. Cornu, "Hidden information detection based on quantized Laplacian distribution," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 1793–1796.
[10] T. H. Thai, R. Cogranne, and F. Retraint, "Statistical model of quantized DCT coefficients : Application in the steganalysis of jsteg algorithm," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 1980–1993, May 2014.
[11] H. Farid, "Detecting hidden messages using higher-order statistical models," in *IEEE International Conference on Image Processing*, vol. 2, 2002, pp. II–905 – II–908.
[12] T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inform. Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.
[13] T. Pevný, J. Fridrich, and A. Ker, "From blind to quantitative steganalysis," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 445–454, April 2012.
[14] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 432–444, April 2012.
[15] T. Pevný and J. Fridrich, "Multiclass detector of current steganographic methods for jpeg format," *Information Forensics and Security, IEEE Transactions on*, vol. 3, no. 4, pp. 635–650, Dec 2008.
[16] V. Sedighi, R. Cogranne, J. Fridrich and "Content-Adaptive Steganography by Minimizing Statistical Detectability," submitted to *Information Forensics and Security, IEEE Transactions on*, 2015
[17] V. Sedighi, J. Fridrich, and R. Cogranne, "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model," in *IS&T/SPIE Electronic Imaging conf.*, vol. 9409, Security, Steganography, and Watermarking of Multimedia Contents, 2015, pp. 94 090H.
[18] R. Cogranne, T. Denemark, and J. Fridrich, "Theoretical model of the FLD ensemble classifier based on hypothesis testing theory," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014, pp. 167–172.
[19] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *Image Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 702–712, 2006.
[20] Y. Su, S. Shan, X. Chen, and W. Gao, "Hierarchical ensemble of global and local classifiers for face recognition," *Image Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1885–1896, 2009.
[21] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *Image Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 407–426, 2008.
[22] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system — the ins and outs of organizing boss," in in *Information Hiding*, ser. LNCS vol.6958, 2011, pp. 59–70.
[23] V. Schwamberger and F. O. Franz, "Simple algorithmic modifications for improving blind steganalysis performance," in *ACM Multimedia & Security*, MM&Sec'10, 2010, pp. 225–230.
[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. John Wiley & Sons, 2012.
[25] T. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.
[26] E. Lehmann and J. Romano, *Testing Statistical Hypotheses, Second Edition*, 3rd ed. Springer, 2005.
[27] R. Cogranne, V. Sedighi, J. Frdrich and T. Pevný, "Is Ensemble Classifier Needed for Steganalysis in High-Dimensional Feature Spaces?," submitted to *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015.
[28] M. Chaumont and S. Kouider, "Steganalysis by ensemble classifiers with boosting by regression, and post-selection of features," in *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1133–1136.
[29] A. D. Ker, "Batch steganography and pooled steganalysis," in *Information Hiding*, ser. LNCS vol. 4437, 2007, pp. 265–281.
[30] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234–239.
[31] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Information Hiding*, ser. LNCS vol. 6387, 2010, pp. 161–177.
[32] T. Filler and J. Fridrich, "Gibbs construction in steganography," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 4, pp. 705–720, Dec 2010.
[33] L. L. Cam, *Asymptotic Methods in Statistical Decision Theory*. New York: springer, 1986.
[34] T. Pevný, "Detecting messages of unknown length," *IS&T/SPIE Electronic Imaging conf.*, vol. 7880, Security, Steganography, and Watermarking of Multimedia Contents, 2014, pp. 78 800T–11.
[35] D. Middleton, *An introduction to statistical communication theory*. McGraw-Hill New York, 1960, vol. 960.
[36] C. W. Helstrom, *Elements of Signal Detection and Estimation*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.
[37] H. Van Trees, *Detection, Estimation, and Modulation Theory*. Wiley, 2004, vol. 1-4.
[38] A. D. Ker and T. Pevný, "A mishmash of methods for mitigating the model mismatch mess," in *IS&T/SPIE Electronic Imaging conf.*, vol. 9028, Security, Steganography, and Watermarking of Multimedia Contents, 2014, pp. 94 028I–14.
[39] B. Baygün and A.O. Hero III, "Optimal simultaneous detection and estimation under a false alarm constraint," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 688 –703, may 1995.
[40] H. Gassmann, I. Deák, and T. Szántai, "Computing multivariate normal probabilities: A new look," *Journal of Computational and Graphical Statistics*, vol. 11, no. 4, pp. 920–949, 2002.
[41] A. Genz and F. Bretz, *Computation of multivariate normal and t probabilities*, ser. Lecture Notes in Statistics, vol. 195, 2009, .
[42] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.
[43] J. Fridrich and J. Kodovskỳ, "Rich models for steganalysis of digital images," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 3, pp. 868 –882, june 2012.
[44] T. Denemark, V. Sedighi, V. Holub, R. Cogranne and J. Fridrich, "Selection-channel-aware rich model for Steganalysis of digital images," in *Information Forensics and Security (WIFS), 2014 IEEE International Workshop on*, 2014, pp. 48–53.
[45] J. Fridrich, T. Pevný, and J. Kodovský, "Statistically undetectable jpeg steganography: Dead ends challenges, and opportunities," in *ACM Multimedia & Security*, MM&Sec'07, 2007, pp. 3–14.
[46] L. Guo, J. Ni, and Y. Q. Shi, "An efficient jpeg steganographic scheme using uniform embedding," in *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, 2012, pp. 169–174.
[47] C. Wang and J. Ni, "An efficient jpeg steganographic scheme based on the block entropy of dct coefficients," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1785–1788.
[48] J. Kodovskỳ and J. Fridrich, "Steganalysis of JPEG images using rich models,"in *IS&T/SPIE Electronic Imaging conf.*, vol. 8303, Security, Steganography, and Watermarking of Multimedia Contents, 2012, pp. 83 030A–13.

APPENDIX A
EXPRESSION FOR THE LIKELIHOOD RATIO AND PROOF OF OPTIMAL PROPERTIES

The goal of this appendix is threefold. First, elements of calculus are provided to establish the form of the LR (12). Then, proofs of Corollary 1, and Proposition 2 are given.

A. Expression of the Likelihood Ratio for Simple Hypothesis and Proof of Proposition 1

Let us recall that after the transformation (9), the problem of detecting hidden data may be described as a choice between

the following hypotheses (11)

$$\begin{cases} \mathcal{H}_0: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_L)\}, \\ \mathcal{H}_1: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\theta}_1, \mathbf{I}_L)\}. \end{cases} \tag{28}$$

Here, the pdf of the multivariate normal distribution of L jointly distributed random variables is denoted

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{-L}{2}} \left|\boldsymbol{\Sigma}\right|^{\frac{-1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{29}$$

with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the expectation and the covariance matrix respectively and $|\boldsymbol{\Sigma}|$ denotes the determinant of matrix $\boldsymbol{\Sigma}$.

From the above definition of tested hypotheses, one can get the following expression for the LR between the hypotheses:

$$\begin{aligned} \frac{p(\mathbf{x}; 0, \mathbf{I}_L)}{p(\mathbf{x}; \boldsymbol{\theta}_1, \mathbf{I}_L)} &= \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta}_1)^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\theta}_1)\right)}{\exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)} \\ &= \exp\left(\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\theta}_1)^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\theta}_1)\right) \\ &= \exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_1 - \frac{1}{2}\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_1\right). \end{aligned} \tag{30}$$

Here, taking any strictly increasing function of the LR $\frac{p(\mathbf{x};0,\mathbf{I}_L)}{p(\mathbf{x};\boldsymbol{\theta}_1,\mathbf{I}_L)}$ does not change the properties of the test, up to the decision threshold; the logarithm of the LR, permits us to write

$$\ln \frac{p(\mathbf{x}; 0, \mathbf{I}_L)}{p(\mathbf{x}; \boldsymbol{\theta}_1, \mathbf{I}_L)} = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_1 - \frac{1}{2}\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_1. \tag{31}$$

Similarly, since $\frac{1}{2}\boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_1$ and $\|\boldsymbol{\theta}_1\|$ are constants whatever the true hypothesis may be, removing the first term and scaling by the second does not change the optimality of the test and immediately yields (12).

From the properties of the multivariate normal distribution, it is immediate to obtain the distribution of the LR

$$\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) = \frac{\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}}}{\|\boldsymbol{\theta}_1\|}, \tag{32}$$

as given in (13), as it is essentially a weighted sum of uncorrelated normally distributed random variables.

The proof of Proposition 1 thus immediately follows from the distribution (13) as it straightforward that:

$$\begin{aligned} &\mathbb{P}_{\mathcal{H}_0}\left(\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) > \tau^{\mathrm{lr}}\right) = 1 - \Phi(\tau^{\mathrm{lr}}) \leq \alpha_0 \\ \Leftrightarrow &1 - \alpha_0 \leq \Phi(\tau^{\mathrm{lr}}) \Leftrightarrow \Phi^{-1}(1-\alpha_0) \leq \tau^{\mathrm{lr}} \end{aligned} \tag{33}$$

with the last equality following from the fact that the standard normal cdf Φ is strictly increasing, while the equality in the inequality ensures the maximization of the power function.

Similarly, the power function of the LRT δ^{lr} immediately follows from distribution (13) as

$$\mathbb{P}_{\mathcal{H}_1}\left(\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) > \tau^{\mathrm{lr}}\right) = 1 - \Phi(\tau^{\mathrm{lr}} - \|\boldsymbol{\theta}_1\|). \tag{34}$$

B. Computing the Threshold that Minimizes P_E, Corollary 1

Using the definition of the proposed LRT, the probability of false alarm (14) and the power function (15), also provided in (33) and (34) respectively, as a function of the decision threshold τ^{lr} can be written:

$$\begin{cases} \alpha_0(\tau^{\mathrm{lr}}) = 1 - \Phi\left(\tau^{\mathrm{lr}}\right), \\ \beta_{\delta^{\mathrm{lr}}}(\tau^{\mathrm{lr}}) = 1 - \Phi\left(\tau^{\mathrm{lr}} - \|\boldsymbol{\theta}_1\|\right). \end{cases} \tag{35}$$

The threshold τ^{lr} that minimizes the total probability of error under equal Bayesian priors, $P_E(\tau^{\mathrm{lr}}) = 1/2\left(\alpha_0(\tau^{\mathrm{lr}}) + 1 - \beta_{\delta^{\mathrm{lr}}}(\tau^{\mathrm{lr}})\right)$, can be obtained by using Equation (35) and differentiating with respect to τ^{lr}. From the definition of $\Phi(x)$ is it immediate that

$$\begin{cases} \frac{d\,\alpha_0(\tau^{\mathrm{lr}})}{d\,\tau^{\mathrm{lr}}} = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{{\tau^{\mathrm{lr}}}^2}{2}\right), \\ \frac{d\,\beta_{\delta^{\mathrm{lr}}}(\tau^{\mathrm{lr}})}{d\,\tau^{\mathrm{lr}}} = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\|\boldsymbol{\theta}_1\|-\tau^{\mathrm{lr}})^2}{2}\right), \end{cases} \tag{36}$$

from which it follows that

$$\frac{d\,P_E(\tau^{\mathrm{lr}})}{d\,\tau^{\mathrm{lr}}} = \frac{1}{\sqrt{2\pi}}\left(\exp\left(\frac{-(\|\boldsymbol{\theta}_1\| - \tau^{\mathrm{lr}})^2}{2}\right) - \exp\left(\frac{-{\tau^{\mathrm{lr}}}^2}{2}\right)\right). \tag{37}$$

Setting the derivative of P_E (37) to zero to find the threshold value τ^{lr} that minimizes P_E leads to

$$\begin{aligned} &\exp\left(-\frac{{\tau^{\mathrm{lr}}}^2}{2}\right) - \exp\left(-\frac{(\|\boldsymbol{\theta}_1\| - \tau^{\mathrm{lr}})^2}{2}\right) = 0, \\ \Leftrightarrow\ &{\tau^{\mathrm{lr}}}^2 = (\|\boldsymbol{\theta}_1\| - \tau^{\mathrm{lr}})^2 \Leftrightarrow 2\tau^{\mathrm{lr}} = \|\boldsymbol{\theta}_1\| \Leftrightarrow \tau^{\mathrm{lr}} = \frac{1}{2}\|\boldsymbol{\theta}_1\|. \end{aligned}$$

which proves the Corollary 1.

C. Proof of Uniformly Most Powerful Property, Proposition 2

Next, to prove that the LRT δ^{lr} is also Uniformly Most Powerful for the case in which the payload R is unknown, it is worth noting that the pdf $p(\mathbf{x}; f(R)\boldsymbol{\theta}_1, \mathbf{I}_L)$, see (29), of the normalized base learners' projections $\widetilde{\mathbf{v}}$ can be written:

$$\begin{aligned} p(\mathbf{x}; f(R)\boldsymbol{\theta}_1, \mathbf{I}_L) = &\exp\left(\Lambda^{\mathrm{lr}}(\mathbf{x}) f(R)\|\boldsymbol{\theta}_1\| - \frac{1}{2}f(R)^2\|\boldsymbol{\theta}_1\|\right) \\ &\times \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right), \end{aligned} \tag{38}$$

which, from the factorization (Fisher–Neyman) Theorem [26, Chapt. 1.9] immediately proves that the proposed LR $\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}})$ is a sufficient statistic for R.

Then, recalling that $f : [0,1] \mapsto \mathbb{R}^+$ is an increasing function, it is immediate to note that $\Lambda^{\mathrm{lr}}(\mathbf{x})$ is also an increasing function of R and hence, it follows from [26, Theorem 3.4.1] that the proposed LRT (8) is a Uniformly Powerful Test (UMP) for solving the hypothesis testing problem (18) with a composite alternative hypothesis $\mathcal{H}_R$.

Rémi Cogranne holds the position of Associate Professor at Troyes University of Technology (UTT). He had received his PhD in Systems Safety and Optimization in 2011 and his engineering degree in computer science and telecommunication in 2008 both from UTT. He has been a visiting scholar at Binghamton University in 2014-2015. During his studies, he took a semester off to teach in a primary school in Ziguinchor, Senegal and studied one semester at Jönköping University, Sweden. His main research interests are in hypothesis testing, steganalysis, steganography, image forensics and statistical image processing.

Jessica Fridrich holds the position of Professor of Electrical and Computer Engineering at Binghamton University (SUNY). She has received her PhD in Systems Science from Binghamton University in 1995 and MS in Applied Mathematics from Czech Technical University in Prague in 1987. Her main interests are in steganography, steganalysis, digital watermarking, and digital image forensic. Dr. Fridrich's research work has been generously supported by the US Air Force and AFOSR. Since 1995, she received 19 research grants totaling over $9 mil for projects on data embedding and steganalysis that lead to more than 160 papers and 7 US patents. Dr. Fridrich is a senior member of IEEE and a member of ACM.