Theoretical model of the FLD ensemble classifier based on hypothesis testing theory

Rémi Cogranne, Tomas Denemark, Jessica Fridrich

Theoretical Model of the FLD Ensemble Classifier Based on Hypothesis Testing Theory

Rémi Cogranne, *Member*, *IEEE*,
ICD - ROSAS - LM2S
Troyes University of Technology - UMR 6281, CNRS
12, rue Marie Curie - B.P. 2060 - 10004 Troyes cedex - France
Email: remi.cogranne@utt.fr

Tomáš Denemark and Jessica Fridrich, *Member*, *IEEE*
Department of Electrical and Computer Engineering
Binghamton University
Binghamton, NY 13902-6000
Email: {tdenema1,fridrich}@binghamton.edu

***Abstract*—The FLD ensemble classifier is a widely used machine learning tool for steganalysis of digital media due to its efficiency when working with high dimensional feature sets. This paper explains how this classifier can be formulated within the framework of optimal detection by using an accurate statistical model of base learners' projections and the hypothesis testing theory. A substantial advantage of this formulation is the ability to theoretically establish the test properties, including the probability of false alarm and the test power, and the flexibility to use other criteria of optimality than the conventional total probability of error. 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

The objective of steganography is to hide a secret message within an innocuous looking cover object, such as a digital image, obtaining thus a stego object that can be sent overtly through an insecure channel. The effort focused on detecting the presence of the hidden message is called steganalysis. Both fields have experienced a rapid development during the previous two decades, see, e.g., [1]. Steganalysis detectors can be built by adopting a statistical model of cover objects [3]–[6] and determining the optimal detection statistic (as referred to in [2]) with respect to a given performance criterion. Alternatively, the detector can be constructed by means of machine learning when representing the cover objects using a suitably chosen feature vector. The FLD[1] ensemble classifier [9] has recently become quite popular among researchers on steganography due to its ability to provide accurate detection and very fast training times for large training data sets and high dimensional feature spaces, which are typically required to detect modern steganographic methods.

For a given cover source, machine learning based steganalysis methods are typically much more powerful than optimal detectors designed from simple models. The theoretical statistical properties of such steganalyzers, however, remain unknown. For example, the false alarm and correct detection probabilities are evaluated empirically on a large set of digital images. While the optimal model-based detectors perform worse in practice, they offer undisputable advantages, such as the ability to guarantee a prescribed false alarm probability and an explicit expression for the detection power.

In the present paper, we leverage the advantages of both approaches by casting the ensemble classifier as an optimal detector derived from an accurate statistical model of the base learner's projections. The theory of hypothesis testing allows us to establish the statistical properties of the detector for a chosen performance criterion, such as computing the highest power one can expect from the ensemble for a prescribed false alarm probability. The proposed methodology is in principle applicable to any ensemble classifier based on linear base learners built on randomly sampled subspaces of the feature space. Numerical simulations as well as experiments on real imagery show the sharpness of the theoretical results and the relevance of the proposed methodology for practical applications.

The present paper is organized as follows. Section IV provides a brief description of the FLD ensemble classifier. Section III presents the proposed statistical model used in this paper, states the steganalysis problem within framework of the hypothesis testing theory, and presents the optimal Likelihood Ratio Test (LRT). The statistical properties of the proposed optimal LRT are also analytically established. Numerical results on a large image database for steganographic methods embedding in both spatial and JPEG domains are presented in Section IV. Finally, Section V summarizes the present work and concludes the paper.

II. FLD Ensemble Classifiers (Background)

We use the following notational conventions in this paper. Matrices will be 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 $\mathcal{X}$.

Modern steganographic methods typically require a high dimensional feature representation of images for accurate detection. The FLD ensemble classifier was originally proposed as an alternative to support vector machines as a scalable machine learning tool that can be efficiently used to build accurate detectors in high dimensional feature spaces and large training data sets. However, the theoretical performance of the

[1]FLD stands for Fisher Linear Discriminant.

ensemble remains unstudied. The present paper focuses on the ensemble classifier as originally proposed in [7] for the BOSS competition [8] and later developed in [9].

Since the FLD is a well-known tool, it is only briefly described in this section. The reader is referred to [10] for a more detailed presentation. Let $\mathbf{f} \in \mathbb{R}^d$ be a (column) vector of d features extracted from one image. Let the training sets of cover and stego image features be matrices of size $d \times N^{\rm trn}$ denoted $\mathbf{C}^{\rm trn} = (\mathbf{c}_1^{\rm trn}, \dots, \mathbf{c}_{N^{\rm trn}}^{\rm trn})$ and $\mathbf{S}^{\rm trn} = (\mathbf{s}_1^{\rm trn}, \dots, \mathbf{s}_{N^{\rm trn}}^{\rm trn})$. The FLD assumes that among 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:

$$\mathcal{C} : \begin{cases} \mathcal{H}_0 \ \text{ if } \ \mathbf{w}^{\rm T}\mathbf{f} - b < 0 \\ \mathcal{H}_1 \ \text{ if } \ \mathbf{w}^{\rm T}\mathbf{f} - b > 0 \end{cases} \tag{1}$$

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

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

Few calculations show that the maximization of the previous Fisher criterion from the training data, $\mathbf{C}^{\rm trn}$ and $\mathbf{S}^{\rm trn}$ leads to the following weighting vector $\mathbf{w}$:

$$\mathbf{w} = \left(\widehat{\boldsymbol{\Sigma}}_c + \widehat{\boldsymbol{\Sigma}}_s\right)^{-1} (\widehat{\boldsymbol{\mu}}_c - \widehat{\boldsymbol{\mu}}_s) \tag{2}$$

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

$$\widehat{\boldsymbol{\Sigma}}_c = \frac{1}{N_{\rm trn} - 1} \sum_{n=1}^{N^{\rm trn}} (\mathbf{c}_n^{\rm trn} - \widehat{\boldsymbol{\mu}}_c)(\mathbf{C}^{\rm trn} - \widehat{\boldsymbol{\mu}}_c)^{\rm T} \ ,$$

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

In principle, the FLD ensemble is a random forest of L base learners implemented as FLDs trained on uniformly randomly selected $d_{\rm sub}$-dimensional subsets $\mathcal{F}_1, \dots, \mathcal{F}_L$ of the feature space. The efficiency of the FLD ensemble classifier comes from fusing the decisions of L such base learners and choosing $d_{\rm sub} << d$, where d is the full feature dimensionality. Let $\mathbf{P}$ be 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 of the corresponding l-th base learner in all remaining elements. Denoting with $\mathbf{b} \in \mathbb{R}^L$ the vector of thresholds of all L base learners (1), the vector of L projections (1) of all base learners can be written as:

$$\mathbf{v} = \mathbf{P}\mathbf{f} - \mathbf{b} \ , \tag{3}$$

where, again, $\mathbf{f} \in \mathbb{R}^d$ is a feature vector to be classified.

In the present paper, the vector $\mathbf{v}$ of base learners' projections is used within the framework of the hypothesis testing theory to design optimal detectors. We remind that because each base learner is trained as a binary classifier, the training requires features from both the cover and the corresponding stego images.

In contrast to [9], we determine the optimal values of $d_{\rm sub}$ and L to match a specified criterion of optimality. In the original formulation of the ensemble, the FLD thresholds were set to minimize the total probability of error under equal Bayesian priors, $P_E = 1/2\,(P_{MD} + P_{FA})$, where P_{MD} and P_{FA} respectively denote the missed detection and false alarm probability (see the formal definition in Section III-A). The methodology proposed in the present paper relies on the Neyman–Pearson criterion of optimality. Hence, during training each detection threshold b is determined to guarantee a prescribed false alarm probability and the parameters $d_{\rm sub}$ and L are chosen as the ones that maximize the power function, see Eq. (12) and (13) in Section III-B. Note that, as in the original version of the FLD ensemble [9], the training set is divided into two subsets, one used for training the FLD base learners, while the second one is used to evaluate the performance of the proposed optimal LR test based on the trained FLD projections. Except when explicitly stated otherwise, all results presented in this paper are obtained with $d_{\rm sub}$ and L determined in this manner.

III. Optimal Binary Detector using Ensemble Classifiers

Let us assume that the vector of base learners' projection $\mathbf{v}$, see Eq. (3), 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 relative payload R and a known embedding method). This constitutes the ideal scenario for the steganalyser as s/he knows the probability distributions 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:

$$\begin{cases} \mathcal{H}_0 : & \{\mathbf{v} \sim \mathcal{P}_{\theta_0}\}\,, \\ \mathcal{H}_1 : & \{\mathbf{v} \sim \mathcal{P}_{\theta_1}\}\,. \end{cases} \tag{4}$$

A statistical test is a mapping $\delta : \mathbb{R}^L \mapsto \{\mathcal{H}_0; \mathcal{H}_1\}$, such that the hypothesis $\mathcal{H}_i$ is accepted if $\delta(\mathbf{v}) = \mathcal{H}_i$ (see [12], [13] for details). The present paper focuses on the Neyman–Pearson bi-criteria approach that minimizes the missed-detection probability for a given false alarm probability. Hence, let:

$$\mathcal{K}_{\alpha_0} = \{\delta : \mathbb{P}_{\mathcal{H}_0}\,(\delta(\mathbf{v}) = \mathcal{H}_1) \leq \alpha_0\}\,, \tag{5}$$

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 $\mathcal{K}_{\alpha_0}$, we need to find a test δ that maximizes the power function defined by the correct detection probability:

$$\beta_\delta = \mathbb{P}_{\mathcal{H}_1}\,(\delta(\mathbf{v}) = \mathcal{H}_1)\,, \tag{6}$$

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 [13, Theorem 3.2.1] that the Most Powerful (MP) test in the class $\mathcal{K}_{\alpha_0}$ (5) is the Likelihood Ratio (LR) test:

$$\delta^{\mathrm{lr}}(\mathbf{v}) = \begin{cases} \mathcal{H}_0 & \text{if } \Lambda^{\mathrm{lr}}(\mathbf{v}) = \dfrac{p_{\theta_1}(\mathbf{v})}{p_{\theta_0}(\mathbf{v})} < \tau^{\mathrm{lr}}, \\ \mathcal{H}_1 & \text{if } \Lambda^{\mathrm{lr}}(\mathbf{v}) = \dfrac{p_{\theta_1}(\mathbf{v})}{p_{\theta_0}(\mathbf{v})} \geq \tau^{\mathrm{lr}}, \end{cases} \tag{7}$$

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 $\mathbb{P}_{\mathcal{H}_0}\left(\Lambda^{\mathrm{lr}}(\mathbf{v}) \geq \tau^{\mathrm{lr}}\right) = \alpha_0$ to ensure that the LR test is in the class $\mathcal{K}_{\alpha_0}$, see Eq. (5).

The choice of the Neyman–Pearson criterion of optimality is justified by practical considerations. When analyzing a large number of digital images the most difficult challenge it to guarantee a low false alarm probability.

A. Statistical Model of Ensemble Classifiers

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, the use of the multivariate normal distribution is supported by invoking Lindeberg's central limit theorem (CLT) [13, Theorem 11.2.5] since the number of features used by each base learner is usually quite large. Using this statistical model, one has $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ 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 expectation and the covariance of base learners' projections under hypothesis $\mathcal{H}_i$, $i = \{0, 1\}$.

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

$$\widetilde{\mathbf{v}} = \boldsymbol{\Sigma}_0^{-1/2}\left(\mathbf{v} - \boldsymbol{\mu}_0\right), \tag{8}$$

where the matrix $\boldsymbol{\Sigma}_0^{-1/2}$ denotes the symmetric matrix satisfying $\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{-1/2} = \boldsymbol{\Sigma}_0^{-1}$ (note that semi-definite positive property of the covariance matrix $\boldsymbol{\Sigma}_0$ ensures uniqueness of $\boldsymbol{\Sigma}_0^{-1/2}$, up to the sign). The affine transformation (8) 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 [13, Chap. 6] and [12, Chap. 4] 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 has been verified on numerical data using BOSS database [8], different embedding methods and different payload. Though not always exact, this assumption is accurate enought in practice, as shown in numerical results provided in Section IV, and especially for small payloads R, which are the focus of the present paper because it is the most difficult case for detection. Unfortunately, due to space limitation we could not include numerical results that especially support this assumption. Note that this assumption, which, roughly speaking, means that stego-embedding "pushes" the expectation of stego-image features in a constant direction is referred to as the "shift hypothesis" and recognized for the first time mentioned in [14].

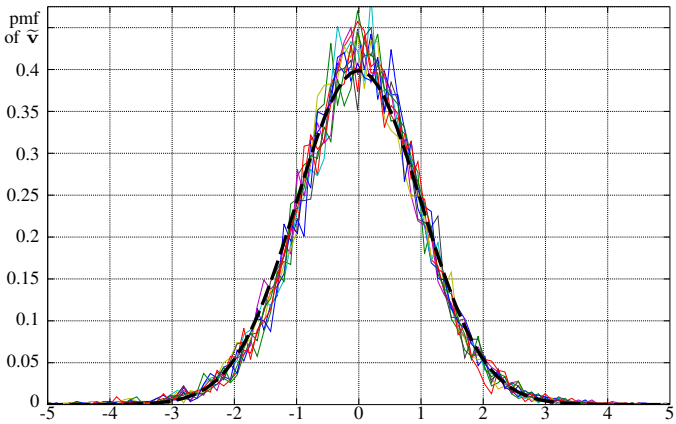


Fig. 1: Comparison between the proposed Gaussian model and the empirical distribution of 10 randomly selected base learners' projections, $\widetilde{\mathbf{v}}$ (8), after normalization.

Let us denote $\boldsymbol{\theta_1} = \boldsymbol{\Sigma}_0^{-1/2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)$. The steganalysis detection problem can be rewritten as a choice between the two following simple hypotheses:

$$\begin{cases} \mathcal{H}_0: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(0, \mathbf{I}_L)\}, \\ \mathcal{H}_1: & \{\widetilde{\mathbf{v}} \sim \mathcal{N}(\boldsymbol{\theta}_1, \mathbf{I}_L)\}. \end{cases} \tag{9}$$

Figure 1 testifies to the accuracy of the proposed multivariate normal model by showing a comparison between the theoretical normal distribution and the empirical distribution of 10 randomly selected normalized base learners' projections $\widetilde{\mathbf{v}}$, see Eq. (8), calculated on one randomly chosen half of images from BOSSbase v1.01 [8] used for testing. The alternative hypothesis for this experiment corresponds to data hidden with WOW [15] at payload $R = 0.05$ bpp (bits per pixel), the feature vector is the 686-dimensional SPAM [16], and the optimal ensemble parameters found were $L = 72$ and $d_{\mathrm{sub}} = 512$.

B. Optimal LR Test and Study of its Statistical Performance

As discussed in the introduction of Section III, the optimal statistical test with a guaranteed false alarm probability and maximal power function for solving the hypothesis testing problem (9) is the LR test defined in Equation (7). In our case, a straightforward calculation shows that the LR between the tested hypotheses can be simplified as:

$$\Lambda^{\mathrm{lr}}(\widetilde{\mathbf{v}}) = \frac{\boldsymbol{\theta}_1^{\mathrm{T}}\widetilde{\mathbf{v}}}{\|\boldsymbol{\theta}_1\|}, \tag{10}$$

where, $\|\boldsymbol{\theta}_1\|^2 = \boldsymbol{\theta}_1^{\mathrm{T}}\boldsymbol{\theta}_1$. 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 Eq. (9), that the

LR $\Lambda^{\text{lr}}(\widetilde{\mathbf{v}})$, Eq. (10), follows:

$$\Lambda^{\text{lr}}(\widetilde{\mathbf{v}}) = \frac{\boldsymbol{\theta}_1^{\text{T}}\widetilde{\mathbf{v}}}{\|\boldsymbol{\theta}_1\|} \sim \begin{cases} \mathcal{N}(0,1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\|\boldsymbol{\theta}_1\|,1) & \text{under } \mathcal{H}_1, \end{cases} \tag{11}$$

From Eq. (11), it is straightforward to establish the statistical properties of the proposed LR test (7) formulated in Proposition 1.

Proposition 1. *For any fixed false alarm probability $\alpha_0 \in (0,1)$ it follows from (11) that the following decision threshold:*

$$\tau^{\text{lr}} = \Phi^{-1}(1-\alpha_0), \tag{12}$$

where Φ and Φ^{-1} denote the standard normal cumulative distribution function (cdf) and its inverse, respectively, guarantees that $\mathbb{P}_{\mathcal{H}_0}\left(\Lambda^{\text{lr}}(\widetilde{\mathbf{v}}) > \tau^{\text{lr}}\right) = \alpha_0$.
From the expression for the threshold τ^{lr}, defined in (12), and the statistical distribution of the LR $\Lambda^{\text{lr}}(\widetilde{\mathbf{v}})$, Equation (11), the power function of the most powerful LR test δ^{lr} is given by:

$$\begin{aligned} \beta_{\delta^{\text{lr}}} = \mathbb{P}_{\mathcal{H}_1}\left(\Lambda^{\text{lr}}(\widetilde{\mathbf{v}}) > \tau^{\text{lr}}\right) &= 1 - \Phi\left(\tau^{\text{lr}} - \|\boldsymbol{\theta}_1\|\right) \\ &= 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \|\boldsymbol{\theta}_1\|\right). \end{aligned} \tag{13}$$

Two essential elements can be deduced from Proposition 1. First, thanks to the normalization of the base learners' projections, see Equation (8), and of the LR $\Lambda^{\text{lr}}(\widetilde{\mathbf{v}})$ through the multiplication by $\|\boldsymbol{\theta}_1\|^{-1}$ (10), the decision threshold only depends on the prescribed false alarm probability and thus guarantees a prescribed false alarm probability. Second, the power function of the optimal LR test only depends on $\|\boldsymbol{\theta}_1\|$, the norm of the expectation under $\mathcal{H}_1$. The expectation $\boldsymbol{\theta}_1$ hence entirely describes the performance of the proposed statistical test.

Figure 2 shows a comparison between the theoretical Gaussian distribution of the LR $\Lambda^{\text{lr}}(\widetilde{\mathbf{v}})$ and the empirical distribution obtained with optimal d_{sub} and L for three different algorithms: WOW [15], S-UNIWARD [17], and HUGO-BD [19] implemented using the Gibbs construction with bounding distortion [20]. The empirical data has been obtained on the testing half of the BOSSbase database.

Remark 1. *It is worth noting that the proposed methodology fundamentally differs from the majority voting rule originally proposed for the FLD ensemble for two main reasons. First, the covariance between the base learners is taken into account. Second, while the majority voting gives the same weight to the base learners, the proposed framework allows giving more importance to base learners that better distinguish the two classes.*
Besides, it should be acknowledged that the computational complexity of the proposed methodology is slightly higher than the one of the original majority voting, though the difference is negligible. In fact, once the FLD projections have been computed, that is vector $\mathbf{v}$, the majority voting consists in only counting either a majority of base learners output '0' or '1'. The proposed methodology requires, in addition, to compute the matrix $\boldsymbol{\Sigma}_0^{-1/2}$ which can be be done efficiently using numerical methods for singular value decomposition (SVD) of matrix $\boldsymbol{\Sigma}_0$. Since the latter matrix is of size $L \times L$, computational complexity is $\mathcal{O}(L^3)$ with L typically smaller than 200.

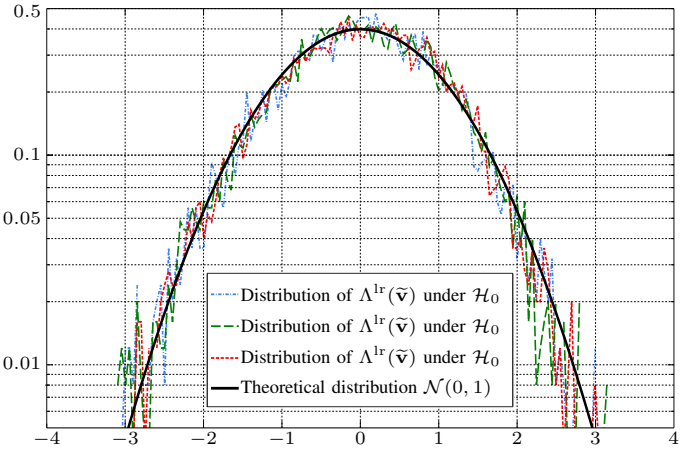


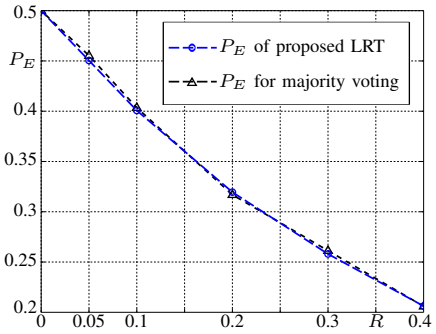
Fig. 2: Comparison between the theoretical normal distribution and the empirical distribution of the proposed LR under $\mathcal{H}_0$ for one half of BOSSbase 1.01 [8] used for testing. The three presented examples correspond to three different alternative hypotheses: WOW, S-UNIWARD, and HUGO-BD, tested with the SRM features.

IV. Numerical Simulations and Results

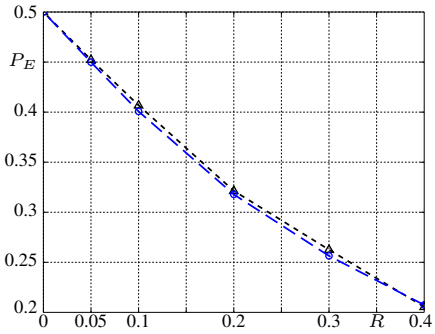
As in the experiments in the previous section, all numerical results presented in this paper are obtained on BOSSbase 1.01. The detection error was always computed by averaging over 10 different random database splits into equally sized subsets. Three spatial domain steganographic algorithms were used: a version of HUGO [19] implemented by minimizing the bounding distortion (HUGO-BD) using the Gibbs construction [20], WOW [15], and S-UNIWARD [17]. The two feature sets used are the second-order SPAM [16] of dimensionality 686 and the Spatial Rich Model (SRM) [18] of dimensionality $34,671$.

Three non side-informed JPEG steganographic algorithms used were nsF5 [21], the Uniform Embedding Distortion (UED) [22], and J-UNIWARD [17]. Three side-informed algorithms were also used: the Perturbed Quantization (PQ) [21], the side-informed version of Entropy-Based Steganography (SI-EBS) [23], and SI-UNIWARD [17]. Four different feature sets were used for steganalysis of JPEGs: the Cartesian-calibrated JPEG Rich Model (CC-JRM) [24] with $22,510$ features, the compact version of JRM referred to as $\mathcal{CF}^\star$ [9], with $7,850$ features, the spatial rich model with one quantization (SRMQ1) [18] of dimensionality $12,753$, and the union of features from SRMQ1 and CC-JRM, referred to as JSRM [24] whose dimensionality is $35,263$.

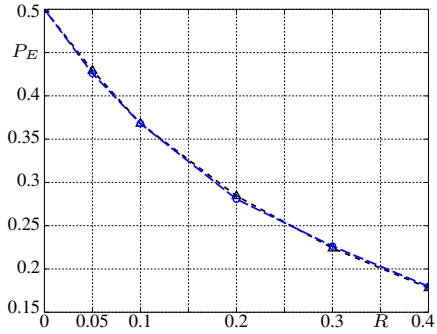
While the main goal of the present paper is to analytically establish the statistical properties of ensemble classifiers within the proposed framework of hypothesis testing, it is also crucial to ensure that the performance of the proposed optimal LR test is comparable to the one obtained with the



(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.

Fig. 3: Comparison between the proposed LR test and the majority vote decision rule for spatial domain steganalysis.

original ensemble classifier with majority voting (EC-MV) as originally proposed in [9]. To this end, the results presented in Figure 3 show a comparison between the EC-MV and the proposed optimal LR test for the SRM feature set and different embedding schemes, WOW in Figure 3a, S-UNIWARD in Figure 3b, and HUGO-BD in Figure 3c. These results were obtained by searching for the optimal parameters d_{sub} and L for each detector. to compare the values of optimal parameters, it is proposed to perform 10 random splits of each feature set reported in Figure 3, at payload $R = 0.2$. The averaged parameters are reported in table I.

Even though Figure 3 shows that the proposed optimal LR test achieves the same performance as the EC-MV, it is not apparent that these two detectors do behave differently with respect to the parameters d_{sub} and L. Figure 4 shows this difference by presenting the performance of both detectors measured as the total probability of error P_E as a function of L for a few fixed values of d_{sub}. The proposed optimal LR test performs much better for small values of L or for small values of d_{sub}. For large values of L and d_{sub} the performance of both detectors becomes almost identical. The results presented in Figure 4 were obtained with the CC-JRM feature set and J-UNIWARD at payload $R = 0.4$ bpnzAC (bits per non-zero AC DCT coefficient). Similar trends have been observed for other feature sets and embedding methods. This phenomenon, together with the difference in the training phase, described in Section , explain the difference observed in practice between the optimal values of parameters d_{sub} and L found for the original EC-MV and the proposed methedology, see table I.

Finally, one of the main goals of the present paper is to use a statistical model of base learners' projections within the framework of hypothesis testing theory to obtain an analytical expression of the proposed test statistical properties. Hence, it is crucial to verify that in practice the theoretically established results accurately hold for real images. Figure 5 shows a comparison between the theoretically established false alarm probability as a function of the decision threshold, $1-\Phi(\tau^{\text{lr}}) = \alpha_0$, see Equation (12), and the empirically measured false alarm probability from the testing set. For brevity, only the results obtained from J-UNIWARD with payload $R = 0.4$ bpnzAC using the JSRM feature set are shown. Similar trends can be found for other embedding methods and feature sets.

The results presented in Figure 5 clearly demonstrate that it is feasible in practice to accurately guarantee a prescribed false alarm probability even for low false alarm probability (typically below $\alpha_0 = 10^{-2}$). We note, however, that a large number of base learners' projections and a high value of d_{sub} make the theoretical results slightly differ from the empirical ones. Note that in practice the optimal values of the parameters d_{sub} and L are smaller than the highest values shown in Figure 5, see table I; especially the value of L which has

Classifier	Ensemble, majority voting (EC-MV)		Optimal LR test (Proposed methodology)	
parameter	d_sub	L	d_sub	L
WOW	2780	95	1550	63
S-UNIWARD	2480	85	1380	61
HUGO-BD	2620	89	1230	75

TABLE I: Comparison between the optimal values of parameters d_{sub} and L found for the original EC-MV classifier and the proposed optimal LR test, average over 10 random splits, with feature sets shown in Figure 3 at payload $R = 0.2$.

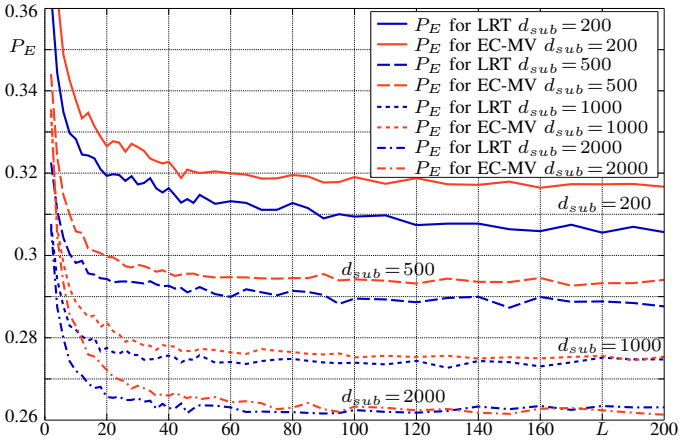


Fig. 4: Comparison between the performance, measured by P_E, of the proposed optimal LRT and the EC-MV detection as a function of L for a few selected values of d_{sub}. The feature set used is CC-JRM and the alternative hypothesis is J-UNIWARD with payload $R = 0.4$ bpnzAC.

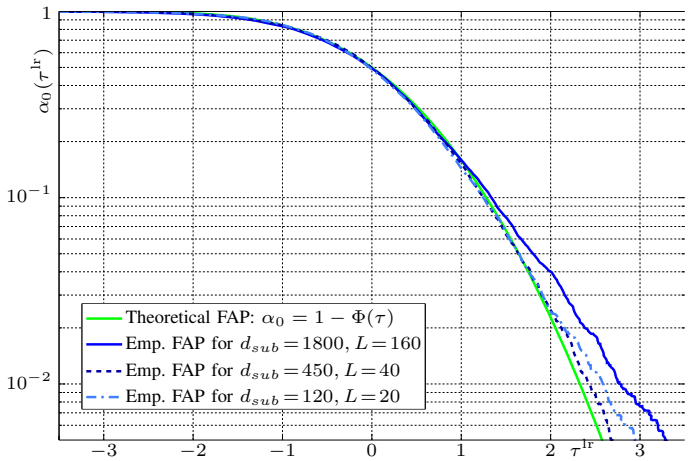


Fig. 5: Comparison between the theoretically established and the empirical probability of false alarm as a function of the decision threshold τ.

the greatest influence on accuracy of results. Those settings have been intentionally chosen to emphasize the limits of the proposed methodology and for readability of Figure 5.

V. Conclusion

This paper proposes a statistical model of base learners' projections in an ensemble classifier, which allows designing an optimal detector with known statistical properties. The main assumptions adopted in this paper are that the base learners' projections follow a multivariate normal distribution and that the covariance matrix remains constant, which is reasonable at least for small payloads. This statistical model is used within the framework of hypothesis testing theory to establish the statistical properties of the optimal LR test. Numerical experiments confirmed the validity of the proposed assumptions that guarantee the accurateness of the theoretically established results.

Acknowledgements

This work was supported by Air Force Office of Scientific 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. The work of Rmi Cogranne is also funded by Troyes University of Technology (UTT) strategic program COLUMBO and STEG-DETECT program for scholar mobility. This research has been done while he was a visiting scholar at Binghamton University.

References

[1] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, 1st ed. Cambridge University Press, 2009.

[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 *Proceedings of the first ACM workshop on Information hiding and multimedia security*, ser. IH&MMSec '13. New York, NY, USA: ACM, 2013, pp. 45–58.

[3] 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.

[4] 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, 2014.

[5] T. H. Thai, F. Retraint, and R. Cogranne, "Statistical detection of data hidden in least significant bits of clipped images," vol. 98, pp. 263 – 274, May 2014.

[6] 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. 1–14, 2014.

[7] J. Fridrich, J. Kodovskỳ, V. Holub, and M. Goljan, "Steganalysis of content-adaptive steganography in spatial domain," in *Information Hiding*, ser. Lecture Notes in Computer Science, T. Filler, T. Pevnỳ, S. Craver, and A. Ker, Eds., vol. 6958. Springer Berlin Heidelberg, 2011, pp. 102–117.

[8] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system — the ins and outs of organizing boss," in *Information Hiding, 13th International Workshop*, ser. Lecture Notes in Computer Science. Prague, Czech Republic: LNCS vol.6958, Springer-Verlag, New York, May 18–20, 2011, pp. 59–70.

[9] J. Kodovskỳ, 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.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. John Wiley & Sons, 2012.

[11] V. Schwamberger and M. O. Franz, "Simple algorithmic modifications for improving blind steganalysis performance," in *Proceedings of the 12th ACM Workshop on Multimedia and Security*, ser. MM&Sec '10. New York, NY, USA: ACM, 2010, pp. 225–230.

[12] T. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

[13] E. Lehmann and J. Romano, *Testing Statistical Hypotheses, Second Edition*, 3rd ed. Springer, 2005.

[14] A. D. Ker, "Batch steganography and pooled steganalysis," in *Information Hiding*, ser. LNCS, vol. 4437, 2007 pp. 265–281.

[15] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, Dec 2012, pp. 234–239.

[16] T. Pevny, 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.

[17] 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.

[18] 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.

[19] T. Pevn, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Information Hiding*, ser. Lecture Notes in Computer Science, R. Bhme, P. Fong, and R. Safavi-Naini, Eds. Springer Berlin / Heidelberg, vol. 6387, pp. 161–177.

[20] 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.

[21] J. Fridrich, T. Pevný, and J. Kodovský, "Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities," in *Proceedings of the 9th Workshop on Multimedia & Security*, ser. MM&Sec'07. New York, NY, USA: ACM, 2007, pp. 3–14.

[22] 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*, Dec 2012, pp. 169–174.

[23] C. Wang and J. Ni, "An efficient JPEG steganographic scheme based on the block entropy of dct coefficients," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 1785–1788.

[24] J. Kodovskỳ and J. Fridrich, "Steganalysis of JPEG images using rich models," vol. 8303, 2012, pp. 83 030A–83 030A–13.