

# Quantitative Structural Steganalysis of Jsteg

Jan Kodovsky and Jessica Fridrich, *Member, IEEE*

**Abstract**—Quantitative steganalysis strives to estimate the change rate defined as the relative number of embedding changes introduced by steganography. In this paper, we propose two new classes of quantitative steganalysis methods for the steganographic algorithm Jsteg. The first class obtains the change-rate estimate using a maximum likelihood estimator equipped with a precover model. While this approach provides better accuracy than existing structural attacks, it becomes computationally intractable with increasing complexity of the cover model. The second class of methods computes the change-rate estimate by minimizing an objective function constructed from a heuristically-formed zero message hypothesis. The advantage of this heuristic approach is a low implementation complexity and modular architecture that allows flexible incorporation of higher-order statistics of DCT coefficients. The proposed methods are experimentally compared with current state-of-the-art methods.

## I. INTRODUCTION

Steganography is the art and science of secret communication between two parties through a public channel. With today's advances in technology and networking and with an ever increasing data traffic over the Internet, the number of communication channels that may be exploited by steganography is rapidly increasing. Progress in steganography usually elicits a corresponding development in steganalysis, whose goal is to establish the existence of steganographic communication. This paper focuses on quantitative steganalysis, which outputs an estimate of the number of embedding changes instead of the binary decision of whether or not the object under investigation contains secret data [1], [2], [3].

Out of more than four hundred data hiding programs available to public today, over a half of them can hide messages in digital images [4]. The first steganographic algorithm for JPEG images, called Jsteg, was developed in 1993 by Upham [5]. Jsteg embeds messages by decompressing the JPEG bit stream to individual quantized DCT coefficients and replacing their Least Significant Bits (LSBs) with message bits. This embedding operation is often being referred to as LSB replacement. In color images, the message is embedded in both the luminance and chrominance components. Jsteg does not embed in coefficients equal to 0 or 1 because too many non-zero coefficients would appear in higher frequencies, which

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org).

The work on this paper was supported by Air Force Office of Scientific Research (AFOSR) under the research grant FA9550-09-1-0147. 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.

Jessica Fridrich is Professor at the Department of Electrical and Computer Engineering, Binghamton University, NY 13902 USA (607-777-6177; fax: 607-777-4464; e-mail: [fridrich@binghamton.edu](mailto:fridrich@binghamton.edu))

would lead to perceptible and statistically detectable artifacts. The first version of Jsteg embedded individual secret data bits sequentially, which turned out to be accurately detectable by the histogram attack [6] capable of estimating the message length as well. An improved version of Jsteg embeds data along a pseudo-random path generated from a secret stego key. In this paper, only this randomized version of Jsteg is considered and will be referred to as Jsteg.

Due to its simplicity and high embedding capacity, the LSB embedding paradigm of Jsteg is the most common embedding principle utilized by creators of steganography software today. Therefore, steganalysis of LSB replacement in DCT domain is of interest to law enforcement and forensic analysts. In particular, it is of interest to detect small payloads that might otherwise go undetected.

The first quantitative attack on Jsteg was described by Zhang and Ping [7]. The authors employed the symmetry of the histogram of DCT coefficients in natural images, which is disturbed by the asymmetric nature of Jsteg embedding operation. Another histogram-based attack was proposed by Yu *et al.* [8]. Currently, the most accurate structural detector of Jsteg is the category attack proposed by Lee *et al.* [9]<sup>1</sup>. Its generalized version [11] has been reported to perform better only on double-compressed images [10]. Even though the category attack outputs a non-negative real value whose expected value on cover images is zero, a direct connection of this test statistic to the message length has not been established. Therefore, this attack is technically not quantitative.

In [12], Westfeld showed that existing quantitative attacks on LSB replacement in the spatial domain<sup>2</sup> can be applied to the array of quantized DCT coefficients to estimate the change rate for Jsteg as well. The explanation of why spatial-domain LSB-replacement detectors work in the frequency domain was provided in [10].

A rather general methodology for constructing quantitative attacks was proposed in [2]. The idea is to model images using a feature vector as in blind steganalysis and capture the relationship between the change in the feature vector to the change rate using regression. According to the reported results, Support Vector Regression (SVR) produced the most accurate quantitative detector of Jsteg today. The disadvantage of this approach is that a large database of images is needed in order to train the regressor. On the other hand, the aforementioned structural detectors operate solely on the image under investigation and do not require any training phase, which might potentially be a problem as the detector might be “overtrained” to a specific source of cover images.

<sup>1</sup>This claim is taken from [10].

<sup>2</sup>For example RS analysis [13], sample pairs analysis [14], or pairs analysis [15].

Apart from feature-based regressors and structural attacks that employ the impact of Jsteg on individual LSB pairs, there exist other approaches for constructing LSB detectors that are based on the Maximum Likelihood (ML) principle [16], [17]. Although ML detectors are more statistically rigorous, their performance is usually weak due to lack of accurate models for cover images<sup>3</sup>. To address this problem, Ker introduced the concept of a precover [18] and proposed a framework that combines the power of structural attacks with the theoretically well-founded ML principle. This idea put existing structural methods on a firmer theoretical ground and outlined possible avenues for their improvement by allowing deviations from model assumptions they were based on.

This paper builds upon the ideas introduced previously, namely the aforementioned concept of precover and least squares steganalysis [18] and introduces a new class of quantitative structural attacks on Jsteg with improved accuracy. The effort begins by investigating the ML approach with precover in Section II, where several new extensions are described. Because the ML approach becomes impractical with increasing complexity of the features, in Section III we explore an alternative direction based on the concept of a zero message hypothesis, a concept introduced already in [13]. Although statistically less rigorous, this heuristic method has some important advantages over the ML approach in terms of low implementation complexity and flexibility. In Section IV, all proposed methods are experimentally evaluated and compared to existing quantitative attacks. The paper is summarized in Section V.

## II. MAXIMUM LIKELIHOOD FRAMEWORK

In this section, we formulate the problem of change-rate estimation in terms of the maximum likelihood principle. The change rate, which will be denoted  $\beta$ , is defined as the relative portion of modified DCT coefficients, with respect to the number of DCT coefficients in the image that are not equal to zero or one. We represent each cover image using a feature vector  $\mathbf{x} \in \mathbb{R}^d$  and assume that it was drawn from some prior distribution  $P_x(\mathbf{x})$ . For a fixed  $\beta$  and  $\mathbf{x}$ , the process of embedding is modeled with the conditional probability  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$ , where  $\mathbf{x}^\beta$  is the feature vector of the stego image affected with change rate  $\beta$ . Here,  $\mathbf{x}^\beta$  is a random variable over the (pseudo) random selection of embedding changes in the cover image  $\mathbf{x}$ . Assuming the cover image and the change rate  $\beta$  are independent of each other,<sup>4</sup> and denoting the stego feature vector  $\mathbf{y}$ ,

$$\begin{aligned} P(\mathbf{y}, \beta) &= \int_{\mathbb{R}^d} P(\mathbf{x}, \mathbf{y}, \beta) d\mathbf{x} = \int_{\mathbb{R}^d} P(\mathbf{y} | \mathbf{x}, \beta) P(\mathbf{x}, \beta) d\mathbf{x} \\ &= \int_{\mathbb{R}^d} P(\mathbf{y} | \mathbf{x}, \beta) P_x(\mathbf{x}) P(\beta) d\mathbf{x}, \end{aligned}$$

which leads to the following ML estimator of  $\beta$ :

<sup>3</sup>This claim appeared in [18].

<sup>4</sup>We make this assumption as it simplifies the analysis while acknowledging that it can be challenged as in reality the sender may adjust the payload size to the cover.

$$\hat{\beta} = \arg \max_{\beta \geq 0} P(\mathbf{y} | \beta) = \arg \max_{\beta \geq 0} \int_{\mathbb{R}^d} P(\mathbf{y} | \mathbf{x}, \beta) P_x(\mathbf{x}) d\mathbf{x}. \quad (1)$$

Practical implementation of this estimator, however, leads to the following complications. While modeling  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$  is usually tractable,<sup>5</sup>  $P_x(\mathbf{x})$  may be rather difficult to obtain. Fortunately, we have an important side-information in the form of the stego image that permits us to eliminate certain image features  $\mathbf{x}$  from our consideration, for example, by realizing that Jsteg embedding preserves the number of DCT coefficients equal to zero or one. The presence of embedding invariants lead the author of [18] to introduce the concept of a precover, which is a hypothetical source the cover is thought of being generated from, given the stego image. The precover is a heuristic and highly feature-dependent concept. It is a means for incorporating into the prior  $P_x(\mathbf{x})$  the cover assumptions and the side-information in the form of the observed stego image with its embedding invariants.

The choice of the features has a crucial impact on the accuracy of the resulting estimator. Obviously, we are interested in feature vectors that are predictably changed by embedding and, at the same time, that can be modeled for typical covers, obtaining thus  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$  and  $P_x(\mathbf{x})$ , respectively. In this section, we consider three different feature vectors  $\mathbf{x}$  and use them, together with the idea of a precover, to construct ML estimators of the form (1). We show that this way, estimators can be built with a comparable or better performance than existing structural attacks.

In the rest of this paper, to avoid repeating the same argument multiple times, all binomial distributions will be approximated with the Gaussian distribution. This is, indeed, justified due to the fact that typical JPEG images contain a large number of DCT coefficients.

### A. Features of Zhang and Ping

Let us consider the following three-dimensional feature vector:

$$\mathbf{x} = [x_1, x_2, x_3] \triangleq [f_0, f_1 - h_1, h_1], \quad (2)$$

where

$$f_0 = \sum_{k>0} h_{2k} + \sum_{k<0} h_{2k+1}, \quad (3)$$

$$f_1 = \sum_{k \geq 0} h_{2k+1} + \sum_{k < 0} h_{2k}, \quad (4)$$

where  $h_i$  is the number of all quantized DCT coefficients in the JPEG image equal to  $i$  (the  $i$ th histogram bin). The quantities (3) and (4) were introduced in [7], together with the Jsteg change-rate estimator<sup>6</sup>

$$\hat{\beta} = \frac{f_1 - f_0}{2h_1}, \quad (5)$$

<sup>5</sup>The features are often selected so that  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$  is easily obtained. In fact, in many structural attacks  $P(\mathbf{x}^\beta | \mathbf{x}, \beta) = \delta(\mathbf{x}^\beta - E[\mathbf{x}^\beta])$ .

<sup>6</sup>In the original publication, the authors estimated the relative payload under the tacit assumption that no matrix embedding is used. In this case, the expected value of the payload is  $2\beta$ , which explains the additional 2 in the denominator of (5).

which we will further refer to as the estimator of Zhang and Ping, or shortly ZP estimator.

We now apply the ML framework for the feature vector defined by (2). Denoting the stego feature

$$\mathbf{x}^\beta = \left[ x_1^\beta, x_2^\beta, x_3^\beta \right],$$

due to the properties of LSB replacement,  $x_1^\beta$  is obtained by drawing from  $x_1$  with probability  $1 - \beta$  and from  $x_2$  with probability  $\beta$ . Thus,  $x_1^\beta$  follows a binomial distribution with mean and variance

$$\mu_1 = (1 - 2\beta)x_1 + \beta C, \quad (6)$$

$$\sigma_1^2 = \beta(1 - \beta)C, \quad (7)$$

where

$$C = x_1 + x_2. \quad (8)$$

Note that the values of  $C$  and  $x_3$  do not change during embedding and  $C + x_3$  is the total number of non-zero coefficients. We will approximate the binomial distribution with a Gaussian  $\varphi(x; \mu_1, \sigma_1^2)$ , where

$$\varphi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}.$$

Because the probability  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$  can be expressed as

$$P(\mathbf{x}^\beta | \mathbf{x}, \beta) = P(x_3^\beta | x_2^\beta, x_1^\beta, \mathbf{x}, \beta) \cdot P(x_2^\beta | x_1^\beta, \mathbf{x}, \beta) \cdot P(x_1^\beta | \mathbf{x}, \beta),$$

due to the embedding invariants

$$\begin{aligned} P(x_3^\beta | x_2^\beta, x_1^\beta, \mathbf{x}, \beta) &= \delta(x_3^\beta = x_3), \\ P(x_2^\beta | x_1^\beta, \mathbf{x}, \beta) &= \delta(x_2^\beta = C - x_1^\beta), \end{aligned}$$

we can write

$$P(\mathbf{x}^\beta | \mathbf{x}, \beta) = \varphi \left( x_1^\beta; \mu_1, \sigma_1^2 \right) \quad (9)$$

for all  $\mathbf{x}^\beta$  that satisfy  $x_3^\beta = x_3$  and  $x_2^\beta = C - x_1^\beta$ , and  $P(\mathbf{x}^\beta | \mathbf{x}, \beta) = 0$  otherwise. We will use this *trick* of reducing the dimensionality of a distribution by incorporating the knowledge of embedding invariants a few more times further in this paper. For this purpose, we introduce the symbol  $\stackrel{ei}{=}$  meaning that the equality holds only when all the embedding invariants are preserved and the probability density is equal to zero otherwise. With this notation, Equation (9) can be rewritten as

$$P(\mathbf{x}^\beta | \mathbf{x}, \beta) \stackrel{ei}{=} \varphi \left( x_1^\beta; \mu_1, \sigma_1^2 \right), \quad (10)$$

without any additional comments on the values of  $x_2^\beta$  and  $x_3^\beta$ .

The prior  $P_x(\mathbf{x})$  will be obtained from a hypothetical precover. Inspired by the equiprobable properties of Ker's precover in [18] and by the reasonable cover assumption  $f_0 = f_1$  used in [7], due to the symmetry of definitions (3) and (4), our precover source will be emitting DCT coefficients from  $f_0$  or  $f_1$  independently and equiprobably. In Appendix A-A, this precover model was verified by testing for normality of  $x_1$  across images. The conclusion is that  $x_1$  is Gaussian only when  $x_1$  is computed from histogram bins  $h_k$ ,  $|k| \geq 3$ . To avoid a significant loss of data and being aware of the model

mismatch, we nevertheless assume  $x_1 \sim N(\mu_2, \sigma_2^2)$ , with  $\mu_2 = \frac{1}{2}(C + x_3)$  and  $\sigma_2^2 = \frac{1}{4}(C + x_3)$ ,

$$P_x(\mathbf{x}) \stackrel{ei}{=} \varphi \left( x_1; \mu_2, \sigma_2^2 \right). \quad (11)$$

Note that in our model for  $P_x(\mathbf{x})$ , we are essentially using the knowledge of the structure of Jsteg embedding by setting  $P_x(\mathbf{x}) = 0$  to all images that cannot be covers for the stego image under investigation.

The ML estimator is obtained by substituting (10) and (11) into (1). (The formula can be simplified because the involved integral can be evaluated analytically). We refrain from including further details of this change-rate estimator because its performance turned out to be essentially identical to the original ZP estimator. The value of this exposition is in the following two comments.

By adopting an additional simplifying assumption that the stego feature  $x_1^\beta$  is equal to its expectation,

$$x_1^\beta = (1 - 2\beta)x_1 + \beta C, \quad (12)$$

the integration in (1) degenerates to a multiplication and the ML estimator reduces to the ZP estimator (5). Second, the quality of estimators that use the histogram symmetry, such as the ZP estimator, may vary significantly depending on whether we count the DC terms when computing the histogram (see Figure 3 in Section IV). Excluding the DC term improves the accuracy because the statistical distribution of DC terms heavily depends on the content and is far from symmetric around zero, which violates the assumption of the detector of Zhang and Ping.

### B. First-Order Statistics

In this section, the histogram truncated to the range  $[-2L, \dots, 2R + 1]$  is used as the feature vector,

$$\mathbf{x} \triangleq [h_{-2L}, \dots, h_{2R+1}]. \quad (13)$$

Since the embedding changes in individual LSB pairs are independent, one can factorize  $P(\mathbf{x}^\beta | \mathbf{x}, \beta)$ :

$$\begin{aligned} P(\mathbf{x}^\beta | \mathbf{x}, \beta) &= P(x_0^\beta | x_0, \beta) \cdot P(x_1^\beta | x_1, \beta) \cdot \\ &\quad \cdot \prod_{k \in I} P(x_{2k}^\beta, x_{2k+1}^\beta | x_{2k}, x_{2k+1}, \beta), \end{aligned} \quad (14)$$

where  $I = \{-L, \dots, R\} \setminus \{0\}$ . From embedding invariants:

$$x_k^\beta = x_k \text{ for } k \in \{0, 1\}, \quad (15)$$

$$x_{2k}^\beta + x_{2k+1}^\beta = x_{2k} + x_{2k+1} \triangleq C_{2k} \text{ for } k \in I, \quad (16)$$

the distribution (14) can be simplified to

$$P(\mathbf{x}^\beta | \mathbf{x}, \beta) \stackrel{ei}{=} \prod_{k \in I} P(x_{2k}^\beta | x_{2k+1}^\beta, x_{2k}, x_{2k+1}, \beta).$$

Approximating the binomial distribution of  $x_{2k}^\beta$  with a Gaussian,  $P(x_{2k}^\beta | x_{2k+1}^\beta, x_{2k}, x_{2k+1}, \beta) = \varphi(x_{2k}^\beta; \mu_{2k}, \sigma_{2k}^2)$  with

$$\begin{aligned} \mu_{2k} &= (1 - 2\beta)x_{2k} + \beta C_{2k}, \\ \sigma_{2k}^2 &= \beta(1 - \beta)C_{2k}, \end{aligned}$$

results in

$$P(\mathbf{x}^\beta | \mathbf{x}, \beta) \stackrel{ei}{=} \prod_{k \in I} \varphi(x_{2k}^\beta; \mu_{2k}, \sigma_{2k}^2). \quad (17)$$

The distribution  $P_x(\mathbf{x})$  is obtained again using the idea of a precover. First, it is assumed that the unquantized DCT coefficients are i.i.d. realizations of a random variable  $\xi$  that follows a generalized Cauchy distribution:

$$g(x) = \frac{p-1}{2s} \left( \left| \frac{x}{s} \right| + 1 \right)^{-p}. \quad (18)$$

This model gave us better results than generalized Gaussian which accords with the results of [8]. The generalized Cauchy distribution is one of many distributions commonly used for modeling the distribution of AC DCT coefficients (see, e.g., [19]). The positive parameters  $p$  and  $s$  were obtained from the stego image using an ML estimator, given the embedding invariants (15) and (16) as integrals of  $g(x)$  over the corresponding regions. Here, we intentionally excluded zeros (the invariant  $\int_{-0.5}^{0.5} g(x)dx = h_0$  was ignored) because the quality of the fit at zero is irrelevant and would only lead to a bias in the other LSB pairs.

The precover is formed by assuming that the histogram bin  $x_{2k}$  is obtained by making  $C_{2k}$  independent draws with probability  $g_{2k} \triangleq P(\xi \in [2k - 0.5, 2k + 0.5] | \xi \in [2k - 0.5, 2k + 1.5])$ , or using the pdf of  $\xi$ :

$$g_{2k} = \left[ \int_{2k-0.5}^{2k+1.5} g(x)dx \right]^{-1} \cdot \int_{2k-0.5}^{2k+0.5} g(x)dx.$$

We approximate the binomial distribution with a Gaussian,  $x_{2k} \sim N(\bar{\mu}_{2k}, \bar{\sigma}_{2k}^2)$ , with

$$\bar{\mu}_{2k} = C_{2k}g_{2k}, \quad (19)$$

$$\bar{\sigma}_{2k}^2 = C_{2k}g_{2k}(1 - g_{2k}). \quad (20)$$

The precover model was verified for LSB pairs  $[x_{2k}, x_{2k+1}]$  with  $|k| \geq 2$  (see Appendix A-B). Again, we adopt the model for all bins to avoid loss of statistical data. Finally,

$$P_x(\mathbf{x}) \stackrel{ei}{=} \prod_{k \in I} \varphi(x_{2k}; \bar{\mu}_{2k}, \bar{\sigma}_{2k}^2). \quad (21)$$

After substituting (17) and (21) into (1), the maximum is found numerically. The computational complexity of this estimator is low because all the involved integrals can be evaluated analytically.

We would like point out the difference between the proposed procedure and the attack of Yu et al. [8]. In [8], authors also use a generalized Cauchy fit as a cover model, but instead of solving the ML equation (1), their estimator is realized using the chi-square test.

### C. Second-Order Statistics

The accuracy of the ML estimator could be further improved by realizing that DCT coefficients are not i.i.d. but exhibit additional dependencies. In this section, we discuss various complications that one encounters when attempting to utilize higher-order statistics of DCT coefficients in the framework based on ML estimation with precover. Even though this

direction is not developed further to construct the estimators, the discussions, terminology, and notation will be used in Section III where an alternative approach to change-rate estimation is pursued. The approach resembles the sample pairs analysis of [14] designed for the spatial domain but is different in due to our usage of the ML framework.

We capture inter-block dependencies among DCT coefficients using an adjacency matrix. Formally, for an image with  $N \times M$  pixels, let us denote the array of DCT coefficients as  $\mathbf{D}_{u,v}(k, l)$ , where  $(k, l)$ ,  $k, l \in \{0, \dots, 7\}$ , is a DCT mode in block  $(u, v)$ ,  $u \in \{0, \dots, [M/8]\}$ ,  $v \in \{0, \dots, [N/8]\}$ . The feature vector is the adjacency matrix  $\mathbf{A} = \{a_{ij}\}$ :

$$a_{ij} = |\{(u, v, k, l) | \mathbf{D}_{u,v}(k, l) = i, \mathbf{D}_{u,v+1}(k, l) = j\}|. \quad (22)$$

Due to the structure of LSB replacement and since Jsteg does not embed into zeros and ones,  $\mathbf{A}$  naturally decomposes into disjoint groups of  $k$  coefficient pairs,  $k \in \{1, 2, 4\}$  (see Figure 1) called  $k$ -nodes. Note that Jsteg embedding can move pairs freely within each node but not among the nodes. The Jsteg embedding transition probabilities for all three  $k$ -node types are shown in Figure 2.

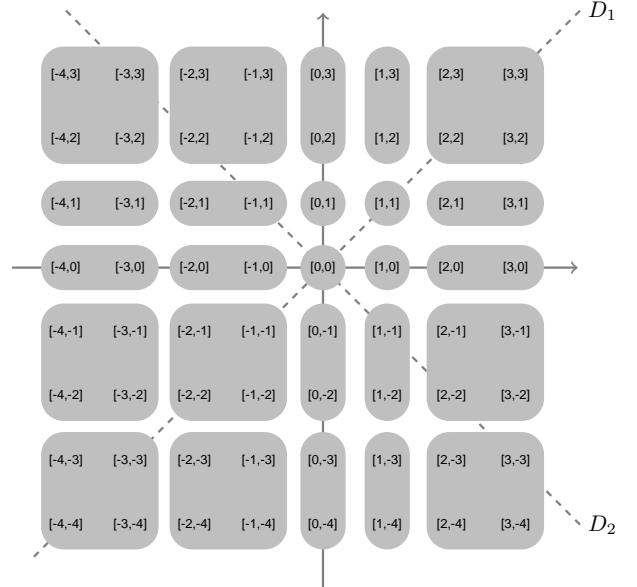


Figure 1. Graphical illustration of the matrix  $\mathbf{A} = \{a_{ij}\}$  defined by (22). Shaded regions represent 1-nodes, 2-nodes, and 4-nodes.

The next step is to derive the model for  $P(\mathbf{A}^\beta | \mathbf{A}, \beta)$  and adopt a model for  $P_x(\mathbf{A})$ . Following Figure 1, the probability  $P(\mathbf{A}^\beta | \mathbf{A}, \beta)$  can be factorized into embedding transition probabilities over individual  $k$ -nodes. The probability for 1-nodes is always equal to 1. The situation for 2-nodes is similar to LSB pairs in a one-dimensional histogram. From the embedding transition probabilities shown in Figure 2 (left), a stego 4-node will follow a multinomial distribution that will be approximated by a multivariate Gaussian distribution. Furthermore, because the sum of occurrences of all four pairs in each 4-node is an embedding invariant, the dimension of the multivariate Gaussian distribution is reduced by one,

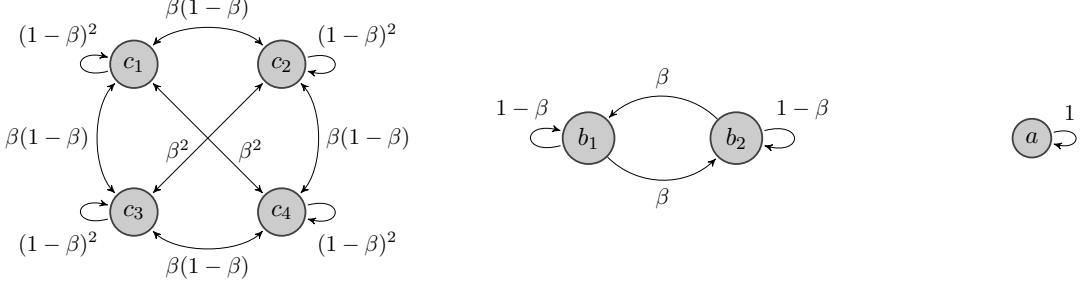


Figure 2. Embedding transition probabilities for all three  $k$ -node types. Left: 4-node  $[c_1, c_2, c_3, c_4]$ , Middle: 2-node  $[b_1, b_2]$ , Right: 1-node  $[a]$ .

resulting in a three-dimensional Gaussian distribution with an appropriate mean and covariance matrix. This way, it is possible to analytically express  $P(\mathbf{A}^\beta | \mathbf{A}, \beta)$  as a product of low-dimensional distributions.

The complications that make this approach to change-rate estimation problematic arise when one attempts to model  $P_x(\mathbf{A})$ . Similarly to the one-dimensional case, the knowledge of embedding invariants can be reflected in  $P_x(\mathbf{A})$  through the precover. After factorizing  $P_x(\mathbf{A})$  into the probabilities over individual  $k$ -nodes, the problem reduces to finding a good parametric model for the (unquantized) cover matrix  $\mathbf{A}$ , given the integrals over the regions corresponding to the individual  $k$ -nodes. However, this is rather difficult because we need to reflect the dependencies between DCT coefficients into the model, otherwise we fundamentally cannot obtain a more accurate detector than in the one-dimensional case. The matrix  $\mathbf{A}$  will exhibit three ridges – one along each axis caused by the fact that coefficients with small absolute value are more frequent than larger ones, and one along the main diagonal, reflecting inter-block dependence of coefficients. Capturing this complicated structure requires using more complex models and estimating more parameters, which increases the complexity of the estimator substantially because this modeling process has to be executed for each analyzed stego image.

Let  $g(\mathbf{A})$  be the statistical model for unquantized coefficient pairs on covers. The precover model assigns the pairs in every  $k$ -node proportionally to the integrals of  $g(\mathbf{A})$  over their corresponding regions. All 2-nodes and 4-nodes will follow a binomial and multinomial distributions, respectively, while 1-nodes will be fully determined by  $g(\mathbf{A})$ . In principle, we can again utilize Gaussian approximations with dimensionality reduced by one thanks to the embedding invariants.

Finally, the model for  $P(\mathbf{A}^\beta | \mathbf{A}, \beta)$  and  $P_x(\mathbf{A})$  can be substituted into the ML estimator (1), carrying out the maximization numerically. Unfortunately, unlike in the one-dimensional case, in every step of the maximization procedure we need to numerically evaluate three-dimensional integrals for all 4-nodes, which further increases the complexity.

To summarize our insight, the complexity of the ML procedure rapidly increases once we start taking into account dependencies among DCT coefficients. Furthermore, the difficulties with modeling  $P_x(\mathbf{A})$ , estimating its parameters from the embedding invariants, and solving (1) motivated us

to investigate alternative strategies for building quantitative steganalysis detectors.

### III. ZERO MESSAGE HYPOTHESIS

In this section, we introduce a simple heuristic change-rate estimating methodology. Even though it was inspired by the ML framework, we leave the probabilistic point of view in favor of a simpler and faster implementation.

#### A. The Proposed Concept

In steganalysis, it is often assumed that the effect of embedding is equal to its expectation [20], [1], [21], [22], which essentially means that the within-image error is ignored. We make this assumption as well:

$$\mathbf{y} = E[\mathbf{x}^\beta] \triangleq \text{Emb}(\mathbf{x}, \beta). \quad (23)$$

Provided  $\text{Emb}$  is invertible, (1) reduces to

$$\hat{\beta} = \arg \max_{\beta \geq 0} P_x(\text{Emb}^{-1}(\mathbf{y}, \beta)), \quad (24)$$

To remove the remaining difficulty of having to model the cover feature statistics  $P_x(\mathbf{x})$ , instead of maximizing the probability (24), we will minimize a penalty  $\mathbf{x}$  evokes. This heuristic measure transforms the problem of modeling  $P_x(\mathbf{x})$  into a much simpler task, namely finding an appropriate penalty function  $z(\mathbf{x}) \geq 0$  that returns zero on cover features and nonzero otherwise. The change-rate estimator (24) becomes:

$$\hat{\beta} = \arg \min_{\beta \geq 0} z(\text{Emb}^{-1}(\mathbf{y}, \beta)). \quad (25)$$

The functional  $z(\mathbf{x})$  is usually obtained from a Zero Message Hypothesis (ZMH) that expresses some key property of covers. The proposed change-rate estimating procedure is summarized as follows:

- 1) Identify a feature vector  $\mathbf{x}$  to be further considered.
- 2) Obtain the inverse embedding operation  $\text{Emb}^{-1}(\mathbf{y}, \beta)$ .
- 3) Identify a ZMH and a penalty functional  $z(\mathbf{x}) \geq 0$ .
- 4) Determine  $\hat{\beta}$  using formula (25).

When designing an estimator, one needs to identify a good feature vector  $\mathbf{x}$ , together with the appropriate ZMH and a penalty functional  $z(\mathbf{x})$  so that the inverse embedding function  $\text{Emb}^{-1}(\mathbf{y}, \beta)$  can be easily constructed. The minimization

in (25) can be carried out either analytically or numerically by implementing a one-dimensional search over  $\beta$ .

Even though this framework is less fundamental when compared to the ML-based methods, its modularity, low computational complexity, and ability to easily incorporate higher-order statistical properties of covers make it appealing. Moreover, it can be used to convert some targeted attacks to quantitative ones. Most importantly, as reported in the experimental Section IV, this approach leads to some of the most accurate change-rate estimators for Jsteg today.

### B. ZP Estimator Revisited

For pedagogical reasons, we first apply the ZMH framework to the feature vector (2). The map Emb

$$\mathbf{x}^\beta = \text{Emb}(\mathbf{x}, \beta) = [(1 - \beta)x_1 + \beta x_2, (1 - \beta)x_2 + \beta x_1, x_3],$$

can be inverted for every  $0 \leq \beta < 1/2$ :

$$\mathbf{x} = \text{Emb}^{-1}(\mathbf{x}^\beta, \beta) = \left[ \frac{x_1^\beta - \beta C}{1 - 2\beta}, \frac{x_2^\beta - \beta C}{1 - 2\beta}, x_3^\beta \right], \quad (26)$$

where  $C$  is an embedding invariant (8)<sup>7</sup>. The ZMH  $f_0 \approx f_1$  can be captured, for example, using

$$z(\mathbf{x}) = (f_0 - f_1)^2 \triangleq (x_1 - x_2 - x_3)^2. \quad (27)$$

After substituting (26) into (27), the minimization w.r.t.  $\beta$  can be carried out analytically and leads exactly to the estimator of Zhang and Ping (5).

### C. Category Attack

We show in this section how to use the proposed framework to convert the category attack to a quantitative one by considering the histogram (13) as a feature vector. For every LSB pair  $[x_{2k}, x_{2k+1}]$ ,  $k \in I$ , the embedding operation can be expressed as

$$\begin{pmatrix} x_{2k}^\beta \\ x_{2k+1}^\beta \end{pmatrix} = \begin{pmatrix} 1 - \beta & \beta \\ \beta & 1 - \beta \end{pmatrix} \begin{pmatrix} x_{2k} \\ x_{2k+1} \end{pmatrix} \quad (28)$$

with its inversion

$$\begin{pmatrix} x_{2k} \\ x_{2k+1} \end{pmatrix} = \frac{1}{1 - 2\beta} \begin{pmatrix} x_{2k}^\beta - \beta C_{2k} \\ x_{2k+1}^\beta - \beta C_{2k} \end{pmatrix}, \quad (29)$$

where the embedding invariants  $C_{2k}$  are defined in (16).

The next step is the formulation of the ZMH. Using a similar notation as in the original publication [9], the category attack uses the function  $S(\mathbf{x}) = (\chi_{\text{shi}}^2 - \chi_{\text{ind}}^2)/(\chi_{\text{shi}}^2 + \chi_{\text{ind}}^2)$ , where

$$\chi_{\text{ind}}^2 = \frac{1}{2} \sum_{k \in I} \frac{(x_{2k} - x_{2k+1})^2}{x_{2k} + x_{2k+1}}, \quad (30)$$

$$\chi_{\text{shi}}^2 = \frac{1}{2} \sum_{k \in I} \frac{(x_{2k} - x_{\Phi(2k-1)})^2}{x_{2k} + x_{2k-1}}, \quad (31)$$

<sup>7</sup>When  $\beta = 1/2$ , the embedding cannot be inverted and  $x_1^\beta = x_2^\beta$ , indicating that a full payload was embedded ( $\hat{\beta} = 1/2$ ). Covers satisfying  $x_1 = x_2$  will lead to a false positive because in this case LSB replacement does not change the feature vector (2). However,  $x_1 = x_2$  is unlikely to hold for natural images.

with  $\Phi(1) = -1$  and  $\Phi(x) = x$ , otherwise. During embedding,  $\chi_{\text{shi}}^2$  increases and  $\chi_{\text{ind}}^2$  decreases. For cover images,  $\chi_{\text{shi}}^2 \approx \chi_{\text{ind}}^2$  or  $S \approx 0$ . This ZMH can be captured using the following penalty function:

$$z_{\text{cat}}(\mathbf{x}) = (S(\mathbf{x}))^2. \quad (32)$$

Minimizing (32) w.r.t.  $\beta$  leads to the quantitative version of the category attack. To the best of our knowledge, the quantitative version of the category attack has not been published yet.

### D. Histogram Symmetry

Keeping the DCT histogram (13) as a feature vector, we can exploit its natural symmetry,<sup>8</sup>  $x_k \approx x_{-k}$ , using the following square penalty function:

$$z_{\text{sym}}(\mathbf{x}) = \sum_{k > 0} w_k (x_k - x_{-k})^2, \quad (33)$$

where the weights  $w_k \geq 0$  are chosen to minimize the variance of the change-rate estimator with penalty function (33). The summation in 33 goes to  $B \triangleq \min\{2L, 2R+1\}$  as we consider only the truncated histogram  $[h_{-2L}, \dots, h_{2R+1}]$ . Since our next steps are essentially identical to Ker's derivation of optimal weighting for least-squares steganalysis [23], [24], we include here only a brief description of the key elements. In particular, the estimator variance is minimized only for stego images with zero payload (covers) under the (precover) assumption that  $x_k$  follows a binomial distribution with size  $x_k + x_{-k}$  and probability  $1/2$ , or  $x_k \sim N(\hat{\mu}_k, \hat{\sigma}_k^2)$ , where  $\hat{\mu}_k = (x_k + x_{-k})/2$  and  $\hat{\sigma}_k^2 = (x_k + x_{-k})/4$ . This precover model is verified in Appendix A-C. The weights  $w_k$  that minimize the variance of  $z_{\text{sym}}(\mathbf{x})$  over cover images are

$$w_k = \frac{1}{x_k + x_{-k}}. \quad (34)$$

We note that the weights (34) are optimal only for cover images (for zero payload) and at least close to optimal for small payloads with no optimality guarantee for larger payloads because in general  $x_k^\beta$  is a poor estimate of  $x_k$ . Nevertheless, being aware of these facts and keeping in mind that the derived weights (34) can be further improved, nothing prevents us from using them in our framework with the penalty function

$$z_{\text{sym}}(\mathbf{x}) = \sum_{k > 0} \frac{(x_k - x_{-k})^2}{x_k + x_{-k}}. \quad (35)$$

### E. Adjacency Attack

In this section, we consider the feature vector formed by the adjacency matrix (22), which captures higher-order statistics between DCT coefficients of the same spatial frequency from two horizontally neighboring  $8 \times 8$  blocks. In Section II-C, we discussed the difficulties when pursuing the ML approach with this feature vector. Apart from a high complexity, the

<sup>8</sup>The DC terms are excluded from histogram calculation in this section as their distribution is not symmetrical.

key difficulty was finding a good cover model  $P_x(\mathbf{A})$ . By switching to estimation using the ZMH, we only need to identify a property of a typical cover matrix  $\mathbf{A}$  that is disturbed by embedding, which is much easier.

First, we quantify the effect of embedding and find the inverse embedding function  $\text{Emb}^{-1}(\mathbf{A}^\beta, \beta)$ . The embedding operation can be studied separately for different types of  $k$ -nodes (follow Figure 2):

- 1-nodes:

$$a^\beta = a,$$

- 2-nodes:

$$\begin{pmatrix} b_1^\beta \\ b_2^\beta \end{pmatrix} = \begin{pmatrix} \gamma & \beta \\ \beta & \gamma \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

- 4-nodes:

$$\begin{pmatrix} c_1^\beta \\ c_2^\beta \\ c_3^\beta \\ c_4^\beta \end{pmatrix} = \begin{pmatrix} \gamma^2 & \beta\gamma & \beta\gamma & \beta^2 \\ \beta\gamma & \gamma^2 & \beta^2 & \beta\gamma \\ \beta\gamma & \beta^2 & \gamma^2 & \beta\gamma \\ \beta^2 & \beta\gamma & \beta\gamma & \gamma^2 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}.$$

Above, we used  $\gamma = 1 - \beta$ . Provided  $0 \leq \beta < 1/2$ , all three embedding functions are linear mappings with a non-singular matrix and thus can be easily inverted, which gives us the inverse mapping  $\text{Emb}^{-1}(\mathbf{A}^\beta, \beta)$ .

To find an appropriate functional  $z(\mathbf{A})$ , we inspect the diagonals denoted  $D_1$  and  $D_2$  in Figure 1. The cover matrix  $\mathbf{A}$  is symmetrical about both diagonals:  $a_{i,j} \approx a_{j,i}$  (the order of DCT coefficients does not matter) and  $a_{i,j} \approx a_{-j,-i}$  (the sign does not matter either).

Because the distribution of DCT coefficients is unimodal with a peak at zero, there are two major ridges in  $\mathbf{A}$  that correspond to the column and row passing through the origin<sup>9</sup>. Note that both ridges are symmetric about both  $D_1$  and  $D_2$ . The matrix  $\mathbf{A}$  has one more, less pronounced, ridge along  $D_1$  due to a small positive correlation between coefficients from neighboring blocks.

Since the symmetry about  $D_2$  is disturbed by embedding (note the asymmetrical placement of  $k$ -nodes w.r.t.  $D_2$  in Figure 1), it will be employed for structural steganalysis through the ZMH framework using the penalty functional expressed again as a weighted sum of square precover deviations:

$$z_{\text{adj}}(\mathbf{A}) = \sum_{i,j=-B}^B \frac{(\bar{a}_{i,j} - \bar{a}_{-i,-j})^2}{\bar{a}_{i,j} + \bar{a}_{-i,-j}}. \quad (36)$$

In (36),  $B = \min\{2L, 2R + 1\}$  determines the size of the largest square submatrix of  $\mathbf{A}$  centered at  $a_{00}$  and  $\bar{a}_{i,j} = a_{i,j} + a_{j,i}$ . The functional (36) is a two-dimensional analogy to the previously introduced optimally weighted one-dimensional penalty function (35). Instead of the histogram symmetry, here we exploit the symmetry of the adjacency matrix along the diagonal  $D_2$ . The optimality of weights in (36) relies on

<sup>9</sup>For higher quality factors, the distribution may not be precisely unimodal, however the coefficients around zero will be more frequent than coefficients with higher absolute values which will again form two ridges in  $\mathbf{A}$ .

the following precover assumption:  $a_{i,j}$  follows a binomial distribution with size  $a_{i,j} + a_{-j,-i}$  and probability 1/2, which is to be simplified as  $a_{i,j} \sim N(\mu_{i,j}, \sigma_{i,j}^2)$  with

$$\begin{aligned} \mu_{i,j} &= \frac{1}{2}(a_{i,j} + a_{-j,-i}), \\ \sigma_{i,j}^2 &= \frac{1}{4}(a_{i,j} + a_{-j,-i}). \end{aligned}$$

This is verified in Appendix A-D. Because the symmetry  $a_{i,j} \approx a_{j,i}$  is preserved under embedding, adding  $a_{i,j}$  and  $a_{j,i}$  to form a new variable  $\bar{a}_{i,j}$  increases the statistical sample and improves the performance.

#### IV. EXPERIMENTAL RESULTS

In this section, we experimentally evaluate all estimators proposed in this paper and compare their performance with current state-of-the-art estimators. The accuracy will be evaluated for change rates ranging from 0 to 0.2. In practice, an image with a negative change rate estimate should be interpreted as cover. To make the results more informative, we do not round negative estimates to zero and instead always perform the minimization in the ZMH framework in the interval  $[-1/2, 1/2]$ .

Because a quantitative steganalysis technique can only estimate the change rate rather than the message length, we used a simulation of Jsteg embedding by directly visiting DCT coefficients (along a pseudo-random path) and flipping a fixed portion  $\beta$  of them. Consequently, the estimation error due to random correlations of the message with the cover elements (e.g., see [3]) is not present in our results.

All experiments were performed on two image databases of small and large images obtained from a mother database of 6,500 JPEG images acquired by 22 different digital cameras at full resolution in a raw format and then converted to grayscale. The size of the images ranged from 1.5 to 6.0 megapixels with a median size of 3.4 megapixels. The first image database of large images was obtained by compressing the images with the JPEG quality factor 75. The second database of small images was obtained by resizing the images before the JPEG compression with quality factor 75. The resizing was carried out using bilinear interpolation so that the smaller side after resizing was 512 pixels (aspect ratio preserved).

Both image datasets were further randomly divided into two equal parts, each consisting of 3,250 images. The first part was used for training of the SVR-based estimator. All remaining methods were then tested on the second part, regardless of the fact of whether or not they required the first half for training. This way, all methods were evaluated on the same set of images, ensuring thus a fair comparison.

In Section IV-A, we present an overall performance comparison by simulating Jsteg on every test image once and analyzing the compound error. In Section IV-B, we analyze the within-image and between-image error for one fixed change rate and two methods.

##### A. Overall Performance

The accuracy of all estimators is reported using the following measures:

- Median absolute error (mAE)

$$\text{median}_i \left\{ \left| \hat{\beta}_i - \beta \right| \right\},$$

- Median bias

$$\text{median}_i \left\{ \hat{\beta}_i - \beta \right\},$$

- Interquartile Range (IQR)

$$\text{iqr}_i \left\{ \hat{\beta}_i \right\}.$$

We additionally calculated the Mean Absolute Error (MAE) but do not report it because both MAE and mAE lead to the same algorithm ranking. We observed that MAE was always proportionally higher than the mAE due to the asymmetric heavy tails. The biggest difference between them was observed for WB and ZP detectors, which indicates that these two estimators have more outliers.

The following quantitative steganalyzers were analyzed: the estimator of Zhang and Ping (5) with and without the DC terms, the histogram-based ML approach described in Section II-B, the Weighted nonsteganographic Borders attack (WB) introduced in [12],<sup>10</sup> the category attack converted to a quantitative one using the ZMH framework, the ZMH-based attack using histograms with  $z_{\text{sym}}(\mathbf{x})$  and using the adjacency matrix with  $z_{\text{adj}}(\mathbf{A})$ , and the quantitative steganalyzer introduced in [2]. Table I conveniently lists all quantitative steganalyzers involved in the test, together with our choices of parameters  $L$ ,  $R$  or  $B = \min\{2L, 2R + 1\}$ . The results for both small and large images are shown in Figure 3.

The following conclusions can be drawn from the experiments:

- 1) The results on both databases are consistent in the sense that the ordering of estimators by their performance is the same on both databases. The estimates on larger images are generally better, which is to be expected due to the square-root law [26].
- 2) In agreement with the results presented in [2], the WB attack has the worst performance.
- 3) The ML approach based on first-order statistics has a comparable performance to the quantitative category attack, but has a larger median bias.
- 4) Among histogram-based attacks, the ZMH method using  $z_{\text{sym}}(\mathbf{x})$  performs the best. In general, the ZMH approach turned out to be very accurate in spite of its heuristic construction.
- 5) Overall, the best experimental performance was achieved using the SVR-based estimator and the ZMH framework with  $z_{\text{adj}}(\mathbf{A})$ . Remarkably, despite the fact that both methods work completely differently, they exhibit very similar accuracy in terms of all three performance measures. The advantage of the proposed ZMH framework over the SVR is that it does not need an expensive training phase – it works solely on an image-by-image basis.

<sup>10</sup>Even though we also tested the Jpairs attack [12], the results are not included in Figure 3 because this method exhibited a markedly worse accuracy compared to the other methods and by including the results in the graphs, their visual clarity would be negatively affected.

## B. Error Analysis

In this section, we investigate the within-image error  $E_W$  and the between-image error  $E_B$  for two selected methods (ZMH-adj and ZP-noDC) on the database of small images. Since the change rate  $\beta$  was fixed to  $\beta = 0.05$ , instead of a real payload, the error caused by random correlations of the message with cover elements was eliminated. This error is unavoidable and the same for every steganalyzer.

To estimate both errors, we used the procedure described in [3], [12], [2]. For every cover image, we define the so-called *cell* as a collection of 200 independent realizations of Jsteg embedding (each with a different stego-key). In order to analyze the tails of the  $E_W$  distribution, one image was randomly selected and we plotted the log-log empirical cumulative distribution function (cdf) [27] of the estimates  $\hat{\beta}$  together with their Gaussian fits – Figure 5 (left column). Overall, the Gaussian fits the collected data well.

In order to study the between-image error  $E_B$ , we remove the normally distributed within-image error component by averaging the estimates over individual cells. The log-log empirical cdf plots of the resulting 3,250 change-rate cell averages are shown in Figure 5 (right column). This time, we provide ML fits with a Gaussian distribution and the Student's *t*-distribution,<sup>11</sup> separately for right and left tails. Unlike  $E_W$ , the between-image error  $E_B$  has thicker tails than the Gaussian fit – they seem to follow the Student's *t*-distribution. These findings are in agreement with previously-published studies of both error sources [3], [2].

## V. CONCLUSIONS

This paper focuses on the steganographic algorithm Jsteg as an archetype of methods that hide message bits in LSBs of quantized DCT coefficients. Two different paradigms for constructing quantitative steganalysis attacks are described and further explored. The first one is a maximum likelihood approach, in which the cover model is derived for each stego image from a hypothetical precover source formed from embedding invariants. While this approach is statistically rigorous, its disadvantage is a rapidly-growing complexity when incorporating higher-order cover models.

In an attempt to address the complexity issue, an alternative heuristic methodology was introduced based on the concept of a zero message hypothesis (ZMH). Instead of modeling the cover source, all that is required is specifying a non-negative penalty function that is approximately zero on covers and that increases with embedding. The change rate is then estimated by minimizing the penalty function using a simple one-dimensional search. This approach is computationally inexpensive and allows easy incorporation of even complicated higher-order properties of covers. The modular architecture also allows reformulating existing targeted (but not quantitative) attacks in terms of the ZMH and, in doing so, convert these attacks to quantitative ones. This was demonstrated on the example of the category attack. Moreover, when the penalty function is a weighted sum of squares of deviations

<sup>11</sup>The ML algorithm was performed w.r.t. both the scale and the degree-of-freedom.

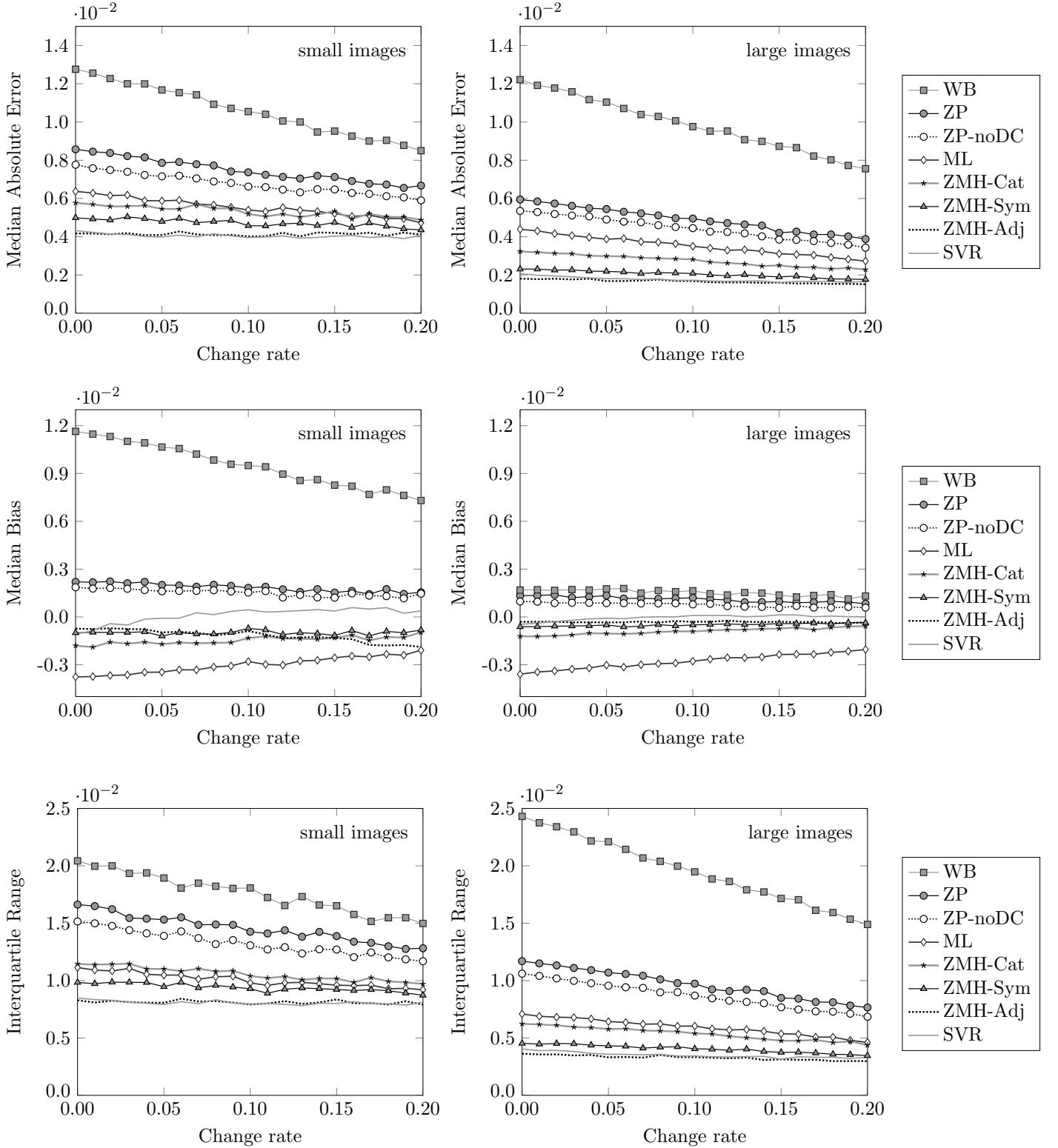


Figure 3. Median absolute error (top), median bias (middle), and interquartile range (bottom) for all methods listed in Table I. The left column corresponds to the database of small images and the right column to large images.

Method	Description
ZP	- Estimator of Zhang and Ping – formula (5). DC terms included.
ZP-noDC	- Estimator of Zhang and Ping – formula (5) when DC terms are excluded.
ML	- Histogram-based ML approach described in Section II-B. $L = 3$ , $R = 2$
WB	- Weighted Nonsteganographic Borders Attack [12]. We used author's code written in R.
ZMH-Cat	- Histogram-based ZMH framework using ZMH $z_{\text{cat}}(\mathbf{x})$ defined by (32). DC terms excluded. $L = R = 4$
ZMH-Sym	- Histogram-based ZMH framework using ZMH $z_{\text{sym}}(\mathbf{x})$ defined by (35). DC terms excluded. $L = R = 4$
ZMH-Adj	- ZMH framework based on the adjacency matrix (22) and $z_{\text{adj}}(\mathbf{A})$ defined in Section III-E. $B = 3$
SVR	- Support vector regression [2] with 548 Cartesian-calibrated Pevný features [25].

Table I  
LIST OF ALL QUANTITATIVE STEGANALYZERS INVOLVED IN EXPERIMENTS.

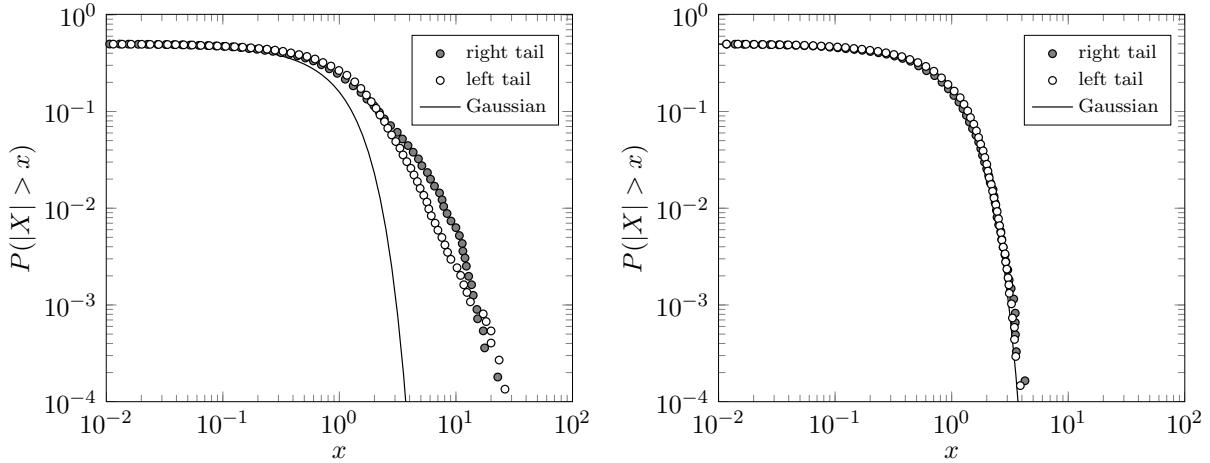


Figure 4. Log-log empirical cdf plot of  $n_1$  (39). Left:  $n_1$  calculated from the whole histogram, Right:  $n_1$  calculated from the histogram after removing the bins  $h_{-4}, h_{-3}, \dots, h_3, h_4$  around zero. Both tails are shown separately. The solid line represents the standard Gaussian distribution.

from the precover model, it is possible to optimize the penalty function by leveraging upon the results derived for optimally-weighted least squares steganalysis [23].

All quantitative steganalysis methods proposed in this paper were experimentally evaluated on a database of 3,250 images and compared to existing attacks. The accuracy of two selected attacks was analyzed in more detail by estimating the between-image and within-image error. The error distributions are compatible with results previously published in the literature. Overall, the heuristic approach based on ZMH provided better results than the maximum likelihood method with a precover. The most accurate estimator was obtained from a ZMH that incorporated the adjacency matrix of DCT coefficients. Its accuracy was comparable to the current state-of-the-art method constructed using support vector regression with a 548-dimensional feature vector. Because the ZMH estimate is only a function of the given stego image, this method offers a much simpler implementation without the need for a potentially expensive training phase.

Among future directions, we mention the possibility to further improve the estimator accuracy by utilizing both intra-block and inter-block dependencies between DCT coefficients by modeling the relationship with Markov chains as in [28].

The fact that such models in their sampled form indeed lead to more accurate blind steganalysis offers reasonable hope that the same quantities and an appropriate ZMH will be useful for quantitative steganalysis of Jsteg as well.

## APPENDIX A VERIFICATION OF PRECOVER ASSUMPTIONS

The purpose of this appendix is to verify four different precover assumptions made in this paper. In all four cases, the task is to confirm whether a sequence of samples  $s^{(i)}$  obtained across different images indexed by  $i$  are realizations of a Gaussian random variable with known mean and variance:

$$s^{(i)} \sim N(\mu^{(i)}, (\sigma^{(i)})^2). \quad (37)$$

Provided hypothesis (37) is a valid statistical descriptor of  $s^{(i)}$ , then  $n \sim N(0, 1)$ , where

$$n = \frac{s^{(i)} - \mu^{(i)}}{\sigma^{(i)}}. \quad (38)$$

This normality assumption was tested on the database of 6,500 (small) cover JPEG images described in Section IV using the Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) test. The former emphasizes the mid-values, while the latter puts

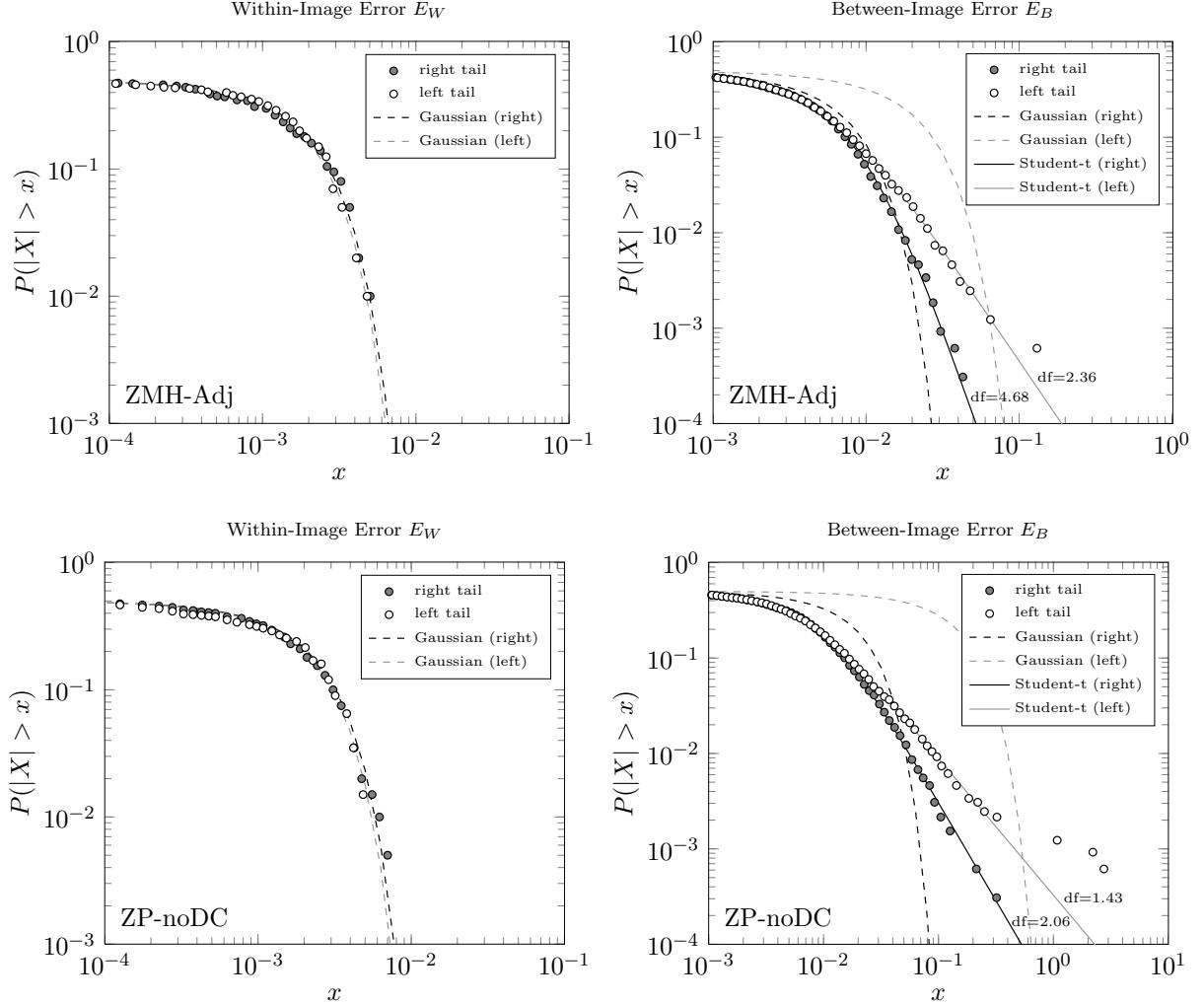


Figure 5. Log-log empirical cdf plots of the within-image error  $E_W$  (left column) and the between-image error  $E_B$  (right column), together with their Gaussian and Student's  $t$  ML fits (for both right and left tails separately). The selected methods were the ZMH-Adj (top), ZP-noDC (bottom). All experiments were conducted on the database of small images.

emphasis on the tails of the distribution. We also observed log-log empirical cdf plots of  $n$  in order to visually verify the Gaussianity of the tails.

#### A. Features of Zhang and Ping

The precover hypothesis states that the feature  $x_1$  follows  $N(\mu_2, \sigma_2^2)$  with  $\mu_2 = \frac{1}{2}(C + x_3)$  and  $\sigma_2^2 = \frac{1}{4}(C + x_3)$  (see Section II-A). Applying the procedure described above, we test whether

$$n_1 = \frac{x_1 - \mu_2}{\sigma_2} \sim N(0, 1) \quad (39)$$

across all 6,500 images<sup>12</sup>. Figure 4 (left) shows a log-log empirical cdf plot of  $n_1$ , together with the reference standard normal distribution. Figure 4 (right) shows the results of the same experiment when the original quantities  $f_0$  and  $f_1$  (and thus  $n_1$ ) were computed only from histogram bins away from zero,  $\dots, h_{-k-2}, h_{-k-1}, h_{k+1}, h_{k+2}, \dots$  with  $k = 4$ . We will

<sup>12</sup>DC terms were excluded when calculating  $f_0$  and  $f_1$  (and thus  $n_1$ ), since they violate histogram symmetry.

call this preprocessing operation as a histogram cut with parameter  $k$ . The results indicate that the precover assumption holds for bins farther from zero, while it does not hold when all histogram bins are included.

In Figure 6, we show the p-values for K-S and A-D normality tests as a function of the histogram cut  $k = 0, \dots, 30$ , together with the 0.05 confidence threshold. Both statistical tests confirm the precover assumption for histogram cuts with  $k \geq 3$ . Note that normality cannot be confirmed for cuts with  $k \geq 25$  due to lack of data. Figure 7 shows the sample means and variances of  $n_1$ , confirming further that (39) follows a standard normal distribution. The see-saw character of the sample mean indicates that natural images contain more positive DCT coefficients than negative ones. This rather strange phenomenon is explained by the presence of downward vertical gradient in blue sky often present in outdoor images.

#### B. First-Order Statistics

In Section II-B, the precover assumption about the cover DCT coefficients in an LSB pair  $[h_{2k}, h_{2k+1}]$  was formulated

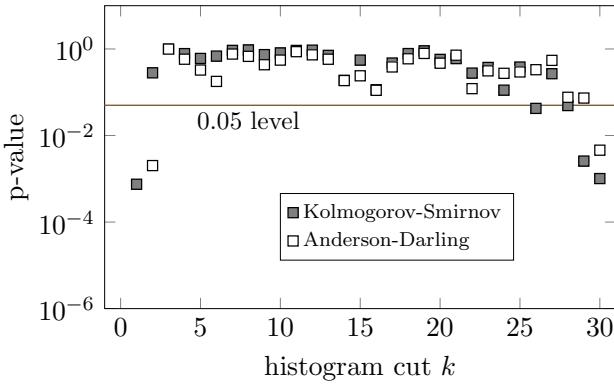


Figure 6. P-values of two normality tests for  $n_1$  shown as a function of histogram cut  $k$ .

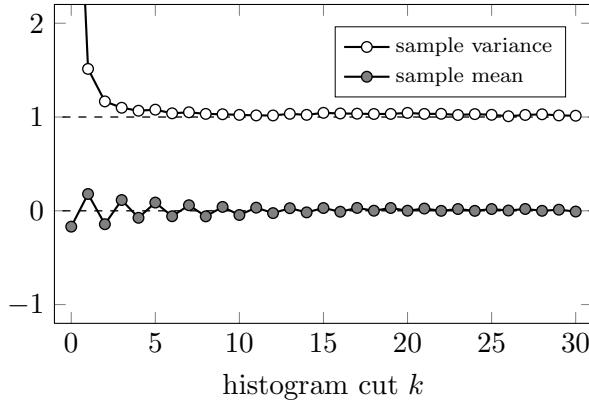


Figure 7. Sample mean and variance of  $n_1$  as a function of the histogram cut  $k$ .

as  $h_{2k} \sim N(\bar{\mu}_{2k}, \bar{\sigma}_{2k}^2)$  with the parameters determined by (19) and (20). Provided this model is a valid statistical description of cover LSB pairs,

$$n_{2k} = \frac{h_{2k} - \bar{\mu}_{2k}}{\bar{\sigma}_{2k}} \sim N(0, 1). \quad (40)$$

This hypothesis was verified using the same methodology as in the previous section and quantified using the K-S and A-D normality tests applied to samples  $n_{2k}$  for  $k \in \{\pm 1, \dots, \pm 15\}$ . The resulting p-values are shown in Figure 8. We conclude that the precover hypothesis holds for pairs  $[h_{2k}, h_{2k+1}]$  with  $|k| \geq 2$ , at least as measured by the K-S test, while the A-D test indicated non-Gaussianity for some values of  $k$ .

### C. Histogram Symmetry

The precover assumption related to the cover-histogram symmetry that appeared in Section III-D is

$$x_k \sim N(\hat{\mu}_k, \hat{\sigma}_k^2).$$

The empirical verification of this hypothesis in terms of the p-values of K-S and A-D tests (applied to its normalized version) are shown in Figure 9. We conclude that even though  $x_k$  has an overall shape similar to Gaussian, it failed the

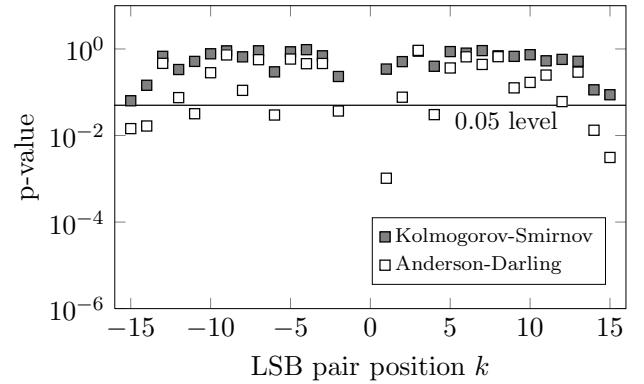


Figure 8. P-values of two normality tests for  $n_{2k}$  as a function of  $k \in \{\pm 1, \dots, \pm 15\}$ .

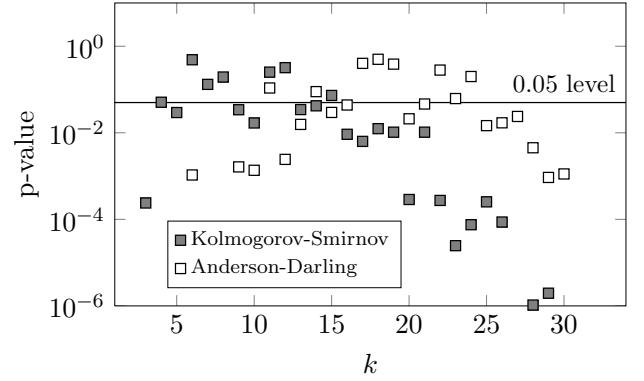


Figure 9. P-values for two normality tests of  $x_k$  (after proper normalization), where  $k = 1, \dots, 30$ .

Gaussianity test. For small values of  $k$ ,  $x_k$  failed the A-D test (Gaussianity of tails), while for larger values of  $k$ ,  $x_k$  failed the K-S test (Gaussianity of the mid-values).

### D. Second-Order Assumption

The precover assumption for adjacency matrix coefficients  $a_{i,j}$  in Section III-E can be stated as  $a_{i,j} \sim N(\mu_{i,j}, \sigma_{i,j}^2)$ , with its normalized version

$$n_{i,j} = \frac{a_{i,j} - \mu_{i,j}}{\sigma_{i,j}} \sim N(0, 1). \quad (41)$$

Again, the normality of  $n_{i,j}$  was verified using the K-S and A-D normality tests. The resulting p-values are shown in Figure 10. While  $n_{i,j}$  passed the K-S test for most of the values from the considered area, it often failed the A-D test. Therefore we cannot confirm Gaussianity, even though the shape of mid-values is very close to the Gaussian distribution (K-S test). Furthermore, we note that the coefficients lying on the diagonal  $D_1 = \{a_{i,i}\}$  did not pass any of the two normality tests. Being aware of the model mismatch, we nevertheless adopt the model for all the coefficients in the considered area, including those lying on  $D_1$ . This is to avoid a significant loss of data as the neighborhood of diagonal  $D_1$  contains a large portion of DCT coefficients due to inter-block correlation.

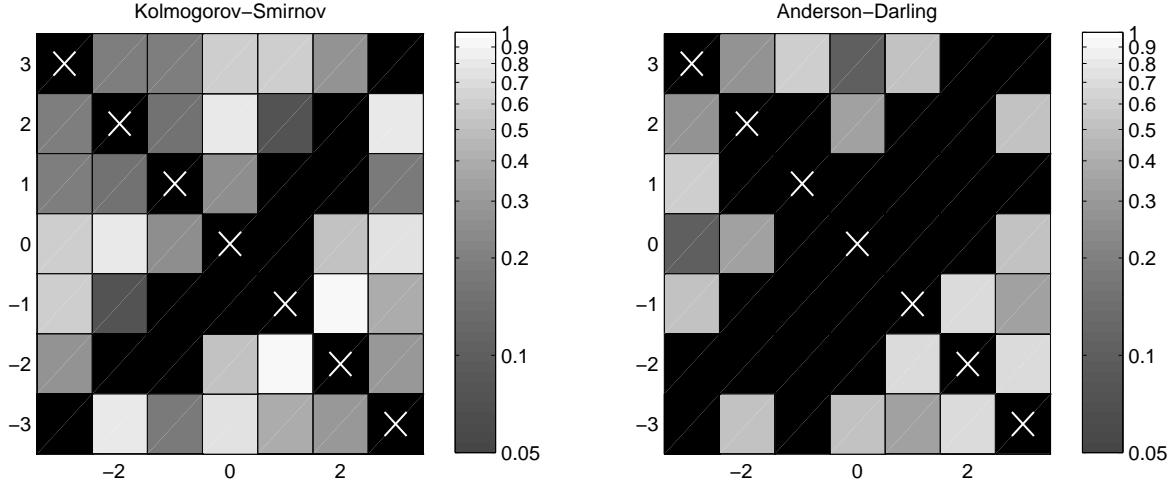


Figure 10. P-values of  $n_{i,j}$  shown as shades of gray for  $i, j \in \{-3, \dots, 3\}$ . Left: K-S, Right: A-D test. Values smaller than 0.05 (in black) correspond to rejecting the hypothesis (41). The crossed coefficients are not part of the precover assumption.

## REFERENCES

- [1] J. Fridrich, M. Goljan, D. Hoga, and D. Soukal, "Quantitative steganalysis of digital images: Estimating the secret message length," *ACM Multimedia Systems Journal*, vol. 9, no. 3, pp. 288–302, 2003.
- [2] T. Pevný, J. Fridrich, and A. D. Ker, "From blind to quantitative steganalysis," in *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XI* (N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, eds.), vol. 7254, (San Jose, CA), pp. 0C 1–0C 14, January 18–21, 2009.
- [3] R. Böhme and A. D. Ker, "A two-factor error model for quantitative steganalysis," in *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VIII* (E. J. Delp and P. W. Wong, eds.), vol. 6072, (San Jose, CA), pp. 59–74, January 16–19, 2006.
- [4] N. F. Johnson and P. Sallee, "Detection of hidden information, covert channels and information flows," in *Wiley Handbook of Science Technology for Homeland Security* (J. G. Voeller, ed.), New York: Wiley & Sons, Inc., April 4, 2008.
- [5] D. Upaham, "Steganographic algorithm JSteg. <http://zooid.org/paul/crypto/jsteg/>"
- [6] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *Information Hiding, 3rd International Workshop* (A. Pfitzmann, ed.), vol. 1768 of Lecture Notes in Computer Science, (Dresden, Germany), pp. 61–75, Springer-Verlag, New York, September 29–October 1, 1999.
- [7] T. Zhang and X. Ping, "A fast and effective steganalytic technique against Jsteg-like algorithms," in *Proceedings of the ACM Symposium on Applied Computing*, (Melbourne, FL), pp. 307–311, March 9–12, 2003.
- [8] X. Yu, Y. Wang, and T. Tan, "On estimation of secret message length in JSteg-like steganography," in *Proceedings of the 17th International Conference on Pattern Recognition* (F. A. P. Petitcolas, ed.), vol. 4, (Cambridge, UK), pp. 673–676, August 23–26, 2004.
- [9] K. Lee, A. Westfeld, and S. Lee, "Category attack for LSB embedding of JPEG images," in *Digital Watermarking, 5th International Workshop* (Y.-Q. Shi, B. Jeon, Y. Shi, and B. Jeon, eds.), vol. 4283 of Lecture Notes in Computer Science, (Jeju Island, Korea), pp. 35–48, Springer-Verlag, Berlin, November 8–10, 2006.
- [10] R. Böhme, "Weighted stego-image steganalysis for JPEG covers," in *Information Hiding, 10th International Workshop* (K. Solanki, K. Sullivan, and U. Madhow, eds.), vol. 5284 of Lecture Notes in Computer Science, (Santa Barbara, CA), pp. 178–194, Springer-Verlag, New York, June 19–21, 2007.
- [11] K. Lee and A. Westfeld, "Generalized category attack – improving histogram-based attack on JPEG LSB embedding," in *Information Hiding, 9th International Workshop* (T. Furun, F. Cayre, G. Doerr, and P. Bas, eds.), vol. 4567 of Lecture Notes in Computer Science, (Saint Malo, France), pp. 378–392, Springer-Verlag, Berlin, June 11–13, 2007.
- [12] A. Westfeld, "Generic adoption of spatial steganalysis to transformed domain," in *Information Hiding, 10th International Workshop* (K. Solanki, K. Sullivan, and U. Madhow, eds.), vol. 5284 of Lecture Notes in Computer Science, (Santa Barbara, CA), pp. 161–177, Springer-Verlag, New York, June 19–21, 2007.
- [13] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and gray-scale images," *IEEE Multimedia, Special Issue on Security*, vol. 8, pp. 22–28, October–December 2001.
- [14] S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via Sample Pairs Analysis," in *Information Hiding, 5th International Workshop* (F. A. P. Petitcolas, ed.), vol. 2578 of Lecture Notes in Computer Science, (Noordwijkerhout, The Netherlands), pp. 355–372, Springer-Verlag, New York, October 7–9, 2002.
- [15] J. Fridrich, M. Goljan, and D. Soukal, "Higher-order statistical steganalysis of palette images," in *Proceedings SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents V* (E. J. Delp and P. W. Wong, eds.), (Santa Clara, CA), pp. 178–190, January 21–24, 2003.
- [16] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, "Detection of hiding in the least significant bit," *IEEE Transactions on Signal Processing*, vol. 52, pp. 3046–3058, 2004.
- [17] M. T. Hogan, N. J. Hurley, G. C. M. Silvestre, F. Balado, and K. M. Whelan, "ML detection of steganography," in *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VII* (E. J. Delp and P. W. Wong, eds.), vol. 5681, (San Jose, CA), January 16–20, 2005.
- [18] A. D. Ker, "A fusion of maximal likelihood and structural steganalysis," in *Information Hiding, 9th International Workshop* (T. Furun, F. Cayre, G. Doerr, and P. Bas, eds.), vol. 4567 of Lecture Notes in Computer Science, (Saint Malo, France), pp. 204–219, Springer-Verlag, Berlin, June 11–13, 2007.
- [19] P. Sallee, "Model-based methods for steganography and steganalysis," *International Journal of Image Graphics*, vol. 5, no. 1, pp. 167–190, 2005.
- [20] S. Dumitrescu, X. Wu, and N. D. Memon, "On steganalysis of random LSB embedding in continuous-tone images," in *Proceedings IEEE, International Conference on Image Processing, ICIP 2002*, (Rochester, NY), pp. 324–339, September 22–25, 2002.
- [21] P. Lu, X. Luo, Q. Tang, and L. Shen, "An improved sample pairs method for detection of LSB embedding," in *Information Hiding, 6th International Workshop* (J. Fridrich, ed.), vol. 3200 of Lecture Notes in Computer Science, (Toronto, Canada), pp. 116–127, Springer-Verlag, Berlin, May 23–25, 2004.

- [22] A. D. Ker, "A general framework for structural analysis of LSB replacement," in *Information Hiding, 7th International Workshop* (M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, eds.), vol. 3727 of Lecture Notes in Computer Science, (Barcelona, Spain), pp. 296–311, Springer-Verlag, Berlin, June 6–8, 2005.
- [23] A. D. Ker, "Optimally weighted least-squares steganalysis," in *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX* (E. J. Delp and P. W. Wong, eds.), vol. 6505, (San Jose, CA), pp. 6 1–6 16, January 29–February 1, 2007.
- [24] A. D. Ker, "Derivation of error distribution in least squares steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 140–148, 2007.
- [25] J. Kodovský and J. Fridrich, "Calibration revisited," in *Proceedings of the 11th ACM Multimedia & Security Workshop* (J. Dittmann, S. Craver, and J. Fridrich, eds.), (Princeton, NJ), September 7–8, 2009.
- [26] T. Filler, A. D. Ker, and J. Fridrich, "The Square Root Law of steganographic capacity for Markov covers," in *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XI* (N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, eds.), vol. 7254, (San Jose, CA), pp. 08 1–08 11, January 18–21, 2009.
- [27] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.
- [28] C. Chen and Y. Shi, "JPEG image steganalysis utilizing both intrablock and interblock correlations," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pp. 3029–3032, May 2008.



**Jan Kodovský** received his M.S. degree in Mathematical Modelling from the Czech Technical University in Prague in 2006. He is currently pursuing the Ph.D. degree at the Electrical and Computer Engineering Department of Binghamton University, New York. His current research interests include steganography, steganalysis, and machine learning.



**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 forensics. 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 \$7.5mil for projects on data embedding and steganalysis that lead to more than 110 papers and 7 US patents. Dr. Fridrich is a member of IEEE and ACM.