# Derivation of Error Distribution in Least Squares Steganalysis

Andrew D. Ker, *Member, IEEE*

***Abstract*—This paper considers the least squares method (LSM) for estimation of the length of payload embedded by least-significant bit replacement in digital images. Errors in this estimate have already been investigated empirically, showing a slight negative bias and substantially heavy tails (extreme outliers). In this paper, (approximations for) the estimator distribution over cover images are derived: this requires analysis of the cover image assumption of the LSM algorithm and a new model for cover images which quantifies deviations from this assumption. The theory explains both the heavy tails and the negative bias in terms of cover-specific observable properties, and suggests improved detectors. It also allows the steganalyst to compute precisely, for the first time, a $p$-value for testing the hypothesis that a hidden payload is present. This is the first derivation of steganalysis estimator performance.**

***Index Terms*—Least-significant bit (LSB) embedding, steganography, structural steganalysis.**

## I. Introduction

STEGANALYSIS is the detection of steganography, and this detection can take a number of forms. Many steganalysis methods are quantitative: not simply a binary decision as to whether an input is a cover or stego object, they estimate the length of the payload (possibly zero). Particularly for LSB replacement steganography in digital images, quantitative detectors seem to present themselves naturally as part of the detection process (see, for example, [1]–[4]).

However, no steganalysis method is perfect so these estimates will be subject to error. In the literature [5], [6], it has become apparent that quantitative detectors for LSB replacement suffer from errors of a pathological type. There are sometimes extreme outliers in the error distribution (the errors appear to be very far from Gaussian) and some estimators, particularly those with the smallest error variance [4], [7], suffer from a small bias. Furthermore, the nature of these errors seems to be highly influenced by the class of image under consideration: the size, local variance, and saturation are shown empirically to be important in [6], but there are likely to be other influences on accuracy. This presents the steganalyst with a difficult problem: given an estimate for the amount of embedded data, how much confidence should they have in it? This goes to the heart of the steganalysis problem. As demonstrated in [8], knowledge of properties of the cover source can make a vast difference to a steganalyst's confidence in their result, but in many applications (such as network monitoring), it probably cannot be assumed that the steganalyst has much information of this sort.

This paper considers a particular quantitative detector for LSB replacement in grayscale images: the least squares method (LSM) variant [9] of sample pairs analysis (SPA) [3]. The aim is to derive its error distribution; it will be possible to do so for one source of error, as long as the detector is modified to remove dependence on a pathological component.

In [6], it is observed that the steganalysis estimator error should be decomposed into two components: within-image error and between-image error. These are separated and their nature investigated empirically for a number of LSB replacement estimators, including the LSM/SPA algorithm. Broadly speaking, the within-image error is due to the content and location of the payload, whereas the between-image error is entirely due to the cover. Although within-image error should not be discounted, it is generally of much smaller magnitude than between-image error, unless the embedded payload is very large, and it always has much smaller, apparently Gaussian, tails (therefore within-image error is not responsible for extreme outliers). Furthermore, when no payload is embedded, there is no within-image error. Therefore, it is sufficient for the steganalyst to know only the between-image error distribution in order to compute a $p$-value for an observed estimate, knowledge of which is a fundamental aim.

In this work then, the focus is only on between-image error, which is the error in the estimator when there is no payload embedded.[1] The aim is to provide a genuine $p$-value for the steganalyst, for testing the hypothesis that no payload is hidden against the alternative that some payload is hidden.

Presenting the steganalysis method now known as WS, [10] is another work which uses some theory to examine steganalysis error. However, it does so in passing (as part of the tuning process for the estimator), only for the less-significant within-image error, and it is not clear that the theory has any connection with experimental practice. Another piece of early work which considers steganalysis error is [11], although this primarily presents heuristic methods for slightly improved detection reliability and is not about the analysis of error per se. Finally, steganalysis error and the factors influencing it can be estimated empirically using large-scale experiments and regression analyses, as in [6] and [12].

Presented here is the first derivation of steganalysis error which does not make unrealistic assumptions about the source

Manuscript received September 19, 2006; revised December 12, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jessica Fridrich.

The author is with the Oxford University Computing Laboratory, Oxford OX1 3QD, U.K. (e-mail: adk@comlab.ox.ac.uk).

Digital Object Identifier 10.1109/TIFS.2007.897265

[1] In some literature (e.g., [1]), this is referred to as detector bias but this term will not be used as it suggests bias in the statistical (i.e., strictly additive) sense, which it is not. Indeed, empirical data in [6] suggest that the between-image error is relative, decreasing with higher embedding rates to zero under maximal embedding.

1556-6013/$25.00 © 2007 IEEE

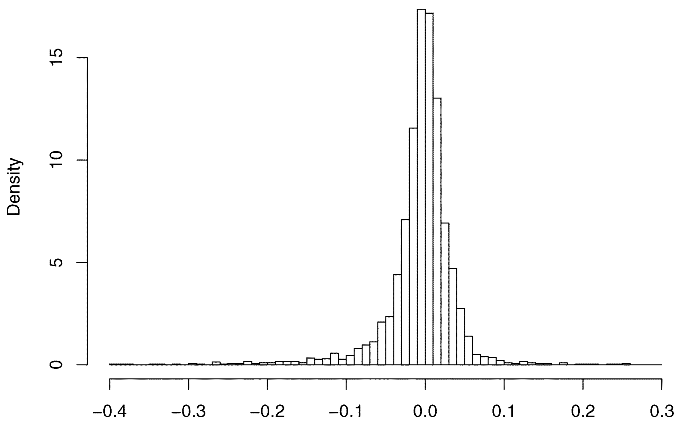

Fig. 1. Histogram of observed LSM/SPA estimates of proportionate length of hidden payload, when none is hidden in 3000 grayscale cover images.

of cover objects, and which accords well with experimental results. But it only applies, at this point, to steganalysis error in cover objects and can only approximate error in stego objects. Despite that limitation, this work has applications both in improved steganalysis and, more speculatively, in adaptive steganography.

As an introductory example, Fig. 1 displays the histogram of the LSM/SPA estimator when applied to 3000 grayscale cover images (no payload is present so the estimator should be around zero). Two features are apparent in this distribution: there is a small negative bias (which turns out to be statistically significant), and a large number of outliers. The distribution does not look Gaussian (it conclusively fails a normality test) and seems somewhat skew. The theory will explain these features in terms of properties of cover images: in fact, the error distribution is approximately Gaussian, but the mean and variance are influenced by image-specific factors so the resulting distribution is a Gaussian mixture.

The structure of this paper is as follows. In Section II, there are some simple mathematics, relating to perturbations in parametric curves of a certain type, which will be a key part of the later derivations. In Section III, the LSM/SPA method is described in just enough detail for the purpose of deriving its error when no payload is embedded. In Section IV, a simple model for cover images is proposed; it explains the errors and is combined with the previous results to derive first- and second-order approximations to the between-image error distribution. It is verified that the second approximation gives a high degree of accuracy in Section V. Applications of this work are briefly presented in Section VI, including a modification of the LSM/SPA method with improved performance. Finally, conclusions are drawn in Section VII.

## II. Small Perturbations in Quadratic Paths

This paper begins with some abstract mathematics. A parametric curve in $\mathbb{R}^m$ will be called a quadratic path if each coordinate is of the form $(s + pt + p^2u)/(1-p)^2$, where $p < 1$ is the parameter and $s, t, u \in \mathbb{R}$ (the reasons for this shape will become apparent later). Its locus will be written in the form

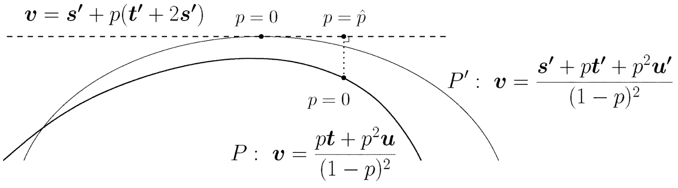
$\boldsymbol{v} = \boldsymbol{s'} + p(\boldsymbol{t'} + 2\boldsymbol{s'})$
$p = 0$
$p = \hat{p}$
$P' : \ \boldsymbol{v} = \dfrac{\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'}}{(1-p)^2}$
$p = 0$
$P : \ \boldsymbol{v} = \dfrac{p\boldsymbol{t} + p^2\boldsymbol{u}}{(1-p)^2}$

Fig. 2. Small perturbations in quadratic paths.

$$\boldsymbol{v} = \frac{\boldsymbol{s} + p\boldsymbol{t} + p^2\boldsymbol{u}}{(1-p)^2}$$

for $p < 1$, where the vectors $\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{u}$ are in $\mathbb{R}^m$. The steganalysis application will be to curves which pass through the origin at $p = 0$ (so $\boldsymbol{s} = \boldsymbol{0}$) and perturbed curves whose coefficients in the numerator are affected by small random vectors. The aim is to estimate $\hat{p}$, the value of the parameter on the perturbed curve which is closest to the origin. First- and second-order approximations for $\hat{p}$ will be found.

Suppose that a quadratic path $P$ passes through the origin at $p = 0$, and that a perturbed path is $P'$. Write $\boldsymbol{v} = (p\boldsymbol{t} + p^2\boldsymbol{u})/(1-p)^2$ for the locus of path $P$ and $\boldsymbol{v} = (\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'})/(1-p)^2$ for $P'$. Approximate $P'$ close to $p = 0$ by its tangent at $p = 0$, which passes through $\boldsymbol{s'}$ and has direction vector $\mathrm{d}\boldsymbol{v}/\mathrm{d}p|_{p=0} = \boldsymbol{t'} + 2\boldsymbol{s'}$. This is closest to the origin at the point whose vector is orthogonal to the direction vector of the tangent (see Fig. 2), so the scalar product $(\boldsymbol{s'} + \hat{p}(\boldsymbol{t'} + 2\boldsymbol{s'})) \cdot (\boldsymbol{t'} + 2\boldsymbol{s'}) = 0$, which occurs when

$$\hat{p} = -\frac{\boldsymbol{s'} \cdot (\boldsymbol{t'} + 2\boldsymbol{s'})}{(\boldsymbol{t'} + 2\boldsymbol{s'}) \cdot (\boldsymbol{t'} + 2\boldsymbol{s'})}.$$

Now identify the perturbations $\boldsymbol{s'} = \boldsymbol{\delta_s}$, $\boldsymbol{t'} = \boldsymbol{t} + \boldsymbol{\delta_t}$, $\boldsymbol{u'} = \boldsymbol{u} + \boldsymbol{\delta_u}$

$$\hat{p} = -\frac{\boldsymbol{\delta_s} \cdot \boldsymbol{t} + \boldsymbol{\delta_s} \cdot (\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s})}{\boldsymbol{t} \cdot \boldsymbol{t} + 2\boldsymbol{t} \cdot (\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s}) + (\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s}) \cdot (\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s})}$$

which, up to first order (i.e., discarding terms whose magnitude is of the order of the square of the perturbations), is

$$\hat{p} \approx -\frac{\boldsymbol{\delta_s} \cdot \boldsymbol{t}}{\boldsymbol{t} \cdot \boldsymbol{t}}. \tag{1}$$

The second-order approximation (discarding terms with magnitude cubic in the perturbations) is obtained by expanding the denominator using the binomial theorem; after some simplification this gives

$$\hat{p} \approx -\frac{\boldsymbol{\delta_s} \cdot \boldsymbol{t}}{\boldsymbol{t} \cdot \boldsymbol{t}} + 2\frac{\big((\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s}) \cdot \boldsymbol{t}\big)(\boldsymbol{\delta_s} \cdot \boldsymbol{t})}{(\boldsymbol{t} \cdot \boldsymbol{t})^2} - \frac{\boldsymbol{\delta_s} \cdot (\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s})}{\boldsymbol{t} \cdot \boldsymbol{t}}. \tag{2}$$

These results will be applied in Section IV.

## III. LSM for Steganalysis of LSB Replacement

Replacement of LSBs in a digital image is an easy (but insecure) method to embed a payload below the visual threshold. The simplest form of embedding traverses the cover in a pseudorandom order and replaces the lowest bits of each pixel value by the payload bits, in which case the number of payload bits is about twice the number of altered cover pixels; more sophisticated methods involving source coding (see, for example, [13]) allow for embedding the same payload with proportionately fewer embedding changes.

The LSM principle [9] leads to a quantitative detector for LSB replacement steganography. More precisely, it is an estimator for twice the number of flipped LSBs in the cover; nonetheless it will be referred to as an estimator for payload size because it was designed for the absence of source coding, but one should bear in mind that it is only the number of flipped LSBs that can truly be estimated. It is based on the sample pairs method of [3], but varies at the final stage when a number of approximate equations are combined to make a single overall estimate. It fits into the structural framework of [4]: a feature vector, which counts the numbers of pixel pairs of certain types, depends in a predictable way on the cover and the number of flipped bits; this relationship can be inverted to see how the same properties of the cover depend on the number of flipped bits and the stego object; finally, there is a model for those properties of cover images. Given a stego image, the payload size is estimated as the value which leads to the closest fit for the cover model.

This estimator will be described in the compact presentation suggested by [4], including only enough detail for subsequent analyses. As in [4], calligraphic letters will be used ($\mathcal{X}$) for sets, uppercase letters ($X$) for random variables, and lowercase letters ($x$) for constants and realizations of random variables. The cover image is (for now) considered constant, and the payload random. Suppose that a digital image consists of a series of $N$ samples with values $s_1, s_2, \ldots, s_N$ in the range $0 \ldots 2M+1$ (typically $M = 127$). A sample pair is a pair of sample locations $(j,k)$ for some $1 \le j \ne k \le N$. Let $\mathcal{P}$ be a set of sample pairs; [3] suggests using all pairs which come from horizontally or vertically adjacent pixels. Then consider some subsets of $\mathcal{P}$

$$\begin{aligned}\mathcal{C}_m &= \left\{(j,k) \in \mathcal{P} \mid \left\lfloor \frac{s_k}{2} \right\rfloor = \left\lfloor \frac{s_j}{2} \right\rfloor + m\right\}\\ \mathcal{E}_m &= \{(j,k) \in \mathcal{P} \mid s_k = s_j + m, \text{ with } s_j \text{ even}\}\\ \mathcal{O}_m &= \{(j,k) \in \mathcal{P} \mid s_k = s_j + m, \text{ with } s_j \text{ odd}\}\end{aligned}$$

for $-M \le m \le M$ in the first case, $-2M \le m \le 2M+1$ in the second, and $-2M+1 \le m \le 2M$ in the third.

Suppose that embedding a payload alters each sample in each pair, independently, with probability $p/2$: this includes the scenario when a payload of length $pN$ (uncorrelated with the cover) is embedded using simple the LSB replacement of a random selection of samples. Alternatively, if a method of source coding is used and the embedding changes are located independently at random, then this same detector will be an estimator for twice the number of embedding changes. The sets $\mathcal{C}_m$ do not involve the LSBs of the pairs, so any pair in $\mathcal{C}_m$ must remain there after LSB replacement. The sets $\mathcal{E}_m$ and $\mathcal{O}_m$ are called trace subsets:[2] each $\mathcal{C}_m$ is partitioned into $\mathcal{E}_{2m}, \mathcal{E}_{2m+1}, \mathcal{O}_{2m-1}, \mathcal{O}_{2m}$, and LSB replacement moves sample pairs among these four trace subsets according to the transition diagram Fig. 3.

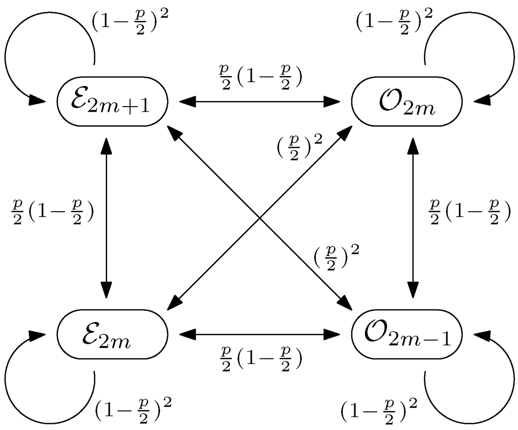

Fig. 3. Transitions between trace subsets when proportionate payload $p$ is embedded by LSB replacement.

Next, count the size of the trace subsets: let $e_m$ (respectively $o_m$) represent the number of sample pairs in $\mathcal{E}_m$ ($\mathcal{O}_m$) before embedding, and the random variable $E'_m$ ($O'_m$) be the number after such a random embedding. Considering Fig. 3, [9] (or, in this notation, [4]) shows

$$\begin{pmatrix} \mathrm{E}[E'_{2m}] \\ \mathrm{E}[O'_{2m-1}] \\ \mathrm{E}[E'_{2m+1}] \\ \mathrm{E}[O'_{2m}] \end{pmatrix} = M\Big(1-\frac{p}{2}, \frac{p}{2}\Big) \begin{pmatrix} e_{2m} \\ o_{2m-1} \\ e_{2m+1} \\ o_{2m} \end{pmatrix} \tag{3}$$

where

$$M(\alpha,\beta) = \begin{pmatrix} \alpha^2 & \alpha\beta & \alpha\beta & \beta^2 \\ \alpha\beta & \alpha^2 & \beta^2 & \alpha\beta \\ \alpha\beta & \beta^2 & \alpha^2 & \alpha\beta \\ \beta^2 & \alpha\beta & \alpha\beta & \alpha^2 \end{pmatrix}.$$

The matrix is invertible as long as $p \ne 1$: the inverse is $1/(1-p)^2 M\big(1-p/2, -p/2\big)$.

Two assumptions are required. First, appealing to the law of large numbers, that the observed realizations $e'_m$ ($o'_m$) of the random variables $E'_m$ ($O'_m$) are close to their expectations

$$e'_m \approx \mathrm{E}[E'_m], \quad o'_m \approx \mathrm{E}[O'_m]. \tag{4}$$

Second, the cover model which drives the estimator

$$e_{2m+1} - o_{2m+1} \approx 0 \tag{5}$$

for $-M \le m < M$. Approximate equations of the form of (5) are termed symmetries in [7]. They are justified by the belief that there will be no correlation between parity structure and pixel difference in continuous-tone images. (The reason for not including also $e_{2m} \approx o_{2m}$ is explained in [4]).

As in [3], it will be very convenient to define $d_m = e_m + o_m$, and $d'_m = e'_m + o'_m$. Putting together (5) with the relevant elements of the inverse of (3), and (4), gives

$$\begin{aligned}0 \approx e_{2m+1} - o_{2m+1} \approx \frac{1}{(1-p)^2}\Big(&e'_{2m+1} - o'_{2m+1}\\ &+ \frac{p}{2}(d'_{2m+2} - d'_{2m} - 2e'_{2m+1} + 2o'_{2m+1})\\ &+ \frac{p^2}{4}(d'_{2m} - d'_{2m+2} + o'_{2m-1} - e'_{2m+3} + e'_{2m+1} - o'_{2m+1})\Big)\end{aligned} \tag{6}$$

[2] Note that these sets are not quite equivalent to those called $\mathcal{X}_m$ and $\mathcal{Y}_m$ used by Dumitrescu *et al.* in [3] or by the original LSM of [9]. Their definition is symmetrical in the sample pairs but introduces an unnecessary special case at $m = 0$. This is explained in [4].

which is an equation for $p$ involving only observations of the stego image. Such an equation can be found for each $m$. The novelty in [9] is to find the value $\hat{p}$ of $p$ which minimizes the sum square error of all of these approximately zero quantities. The mechanics of how such a $p$ may be determined will not be included here, as this may be found already in [9][3] and is not relevant to subsequent analysis.

Two assumptions were made: (4) and (5). The former is responsible for within-image error, the latter for between-image error. As stated in Section I, the analysis in this paper will disregard the within-image error and concentrates only on the between-image error, looking at the steganalysis estimation when no payload is hidden. In that case, $e'_m = e_m$ and $o'_m = o_m$, (4) is redundant, and (6) becomes

$$\frac{s'_m + pt'_m + p^2u'_m}{(1-p)^2} = 0 \tag{7}$$

where

$$\begin{aligned} s'_m &= e_{2m+1} - o_{2m+1} \\ t'_m &= \frac{1}{2}(d_{2m+2} - d_{2m}) - (e_{2m+1} - o_{2m+1}) \\ u'_m &= \frac{1}{4}(d_{2m} - d_{2m+2} + o_{2m-1} - e_{2m+3} + e_{2m+1} - o_{2m+1}). \end{aligned} \tag{8}$$

Therefore, the least squares estimator can be given a geometric interpretation by

$$\hat{p} = \arg\min_p \left\| \frac{\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'}}{(1-p)^2} \right\|$$

where $\boldsymbol{s'}$ (respectively, $\boldsymbol{t'}$, $\boldsymbol{u'}$) are vectors whose entries are each $s'_m$ ($t'_m$, $u'_m$) for $-M \leq m \leq M$, and $\|\cdot\|$ represents the $L^2$-norm. $\hat{p}$ is the parameter where the quadratic path $\boldsymbol{v} = (\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'})/(1-p)^2$ is closest to the origin.

## IV. Derivation of Between-Image Error

Now approximations will be derived for the distribution of the between-image error when the LSM algorithm is used. The key component is a model for natural images which explains deviations from the cover assumption (5). Note that the cover image is no longer considered constant, as it was in Section III, but subject to random "error." But the notation will not change, so the reader is warned that some lowercase letters are now random variables.

### *A. Model for Symmetry Deviation*

The assumption $e_m \approx o_m$ is natural, but it does not hold precisely in images. A model is sought for cover images which explains the deviations from exact equality. It is desirable for the model to be as gentle as possible (and, crucially, not parametric) so that it is not too dependent on the image source for its accuracy.

*1) Model:* [4] Consider the set of all sample pairs in a natural image. Of all those pairs whose values differ by $m$, the first value in each pair is even or odd with probability 1/2, independent of other pairs.

That is, the difference histogram (the frequency of differences of adjacent pixels) is assumed fixed, and the parity of the first pixel in each pair is uniformly random. Of course, this does not reflect the construction of images, but nonetheless it represents a plausible hypothesis about parity structure in a continuous-tone image. An isolated test of this model will be made in Section V-B: it will be seen to be quite accurate for $|m| > 3$, marginally so when $|m| = 3$, and not accurate for $|m| \leq 2$. In this work, nothing more will be done other than restricting the analysis by altering the LSM detector to avoid using the assumption (5) in cases where this model does not fit well.

It would be possible to make stronger assumptions about cover images (e.g., to model the shape of the difference histogram; it is common in the literature to use a generalized Gaussian distribution). However, this temptation is to be resisted for now: the more imposing the assumption is, the less widely applicable it will be.

Given the model, $d_m$ are constants but the $e_m$ are binomial random variables; $e_m$ then determines $o_m$. Making the Gaussian approximation $e_m \sim \mathrm{Bi}(d_m, 1/2) \approx \mathrm{N}(d_m/2, d_m/4)$ (valid as long as $d_m$ is at least about 10; see, for example, [14]) leads to

$$e_m - o_m = 2e_m - d_m \sim \mathrm{N}(0, d_m).$$

For steganalysis by the LSM algorithm, it is only necessary to apply the model for odd values of $m$. It will be convenient to write $e_{2m+1} - o_{2m+1} = \varepsilon_m = \sqrt{d_{2m+1}}Z_m$, so that the $Z_m$ are iid standard Gaussian random variables encompassing all of the randomness in deviations from the exact equations $e_{2m+1} = o_{2m+1}$.

### *B. Distribution of Between-Image Error*

Consider the geometric presentation of the LSM estimator. It is the parameter which places $\boldsymbol{v} = (\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'})/(1-p)^2$ closest to the origin. Write $\boldsymbol{s'} = \boldsymbol{s} + \boldsymbol{\delta_s}$, and so on, where $\boldsymbol{s}$, $\boldsymbol{t}$, and $\boldsymbol{u}$ are the values of $\boldsymbol{s'}$, $\boldsymbol{t'}$, $\boldsymbol{u'}$ when (5) holds exactly, and the perturbations $\boldsymbol{\delta_s}$, $\boldsymbol{\delta_t}$, $\boldsymbol{\delta_u}$ are due to $\boldsymbol{\varepsilon}$. Using (8), the following is derived:

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{0} & \boldsymbol{\delta_s} &= \boldsymbol{\varepsilon} \\ t_m &= \frac{1}{2}(d_{2m+2} - d_{2m}) & \boldsymbol{\delta_t} &= -\boldsymbol{\varepsilon} \end{aligned}$$

(there is no need to know $\boldsymbol{u}$ or $\boldsymbol{\delta_u}$). The first-order approximation (1) gives

$$\hat{p} \approx -\frac{\boldsymbol{\varepsilon}\cdot\boldsymbol{t}}{\boldsymbol{t}\cdot\boldsymbol{t}} = \frac{-2\sum_m (d_{2m+2} - d_{2m})\sqrt{d_{2m+1}}Z_m}{\sum_m (d_{2m+2} - d_{2m})^2}$$

which implies that $\hat{p}$ has a Gaussian distribution $\hat{p} \sim \mathrm{N}\big(\mu_1, v(\boldsymbol{d})\big)$, where

$$\mu_1 = 0, \quad v(\boldsymbol{d}) = \frac{4\sum_m (d_{2m+2} - d_{2m})^2 d_{2m+1}}{\left(\sum_m (d_{2m+2} - d_{2m})^2\right)^2}. \tag{9}$$

Note that if the relative shape of $\boldsymbol{d}$ is fixed and the size of cover $N$ varies, $d_m$ is $O(N)$, implying that $v(\boldsymbol{d}) = O(N^{-1})$.

The second-order approximation leads to a more complicated distribution. Equation (2) is simplified because, here, $\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s} = \boldsymbol{\varepsilon}$. Write $X = \boldsymbol{\varepsilon}\cdot\boldsymbol{t}/\boldsymbol{t}\cdot\boldsymbol{t}$ and $Y = \boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}/\boldsymbol{t}\cdot\boldsymbol{t}$. Then, (2) reduces to

$$\hat{p} \approx -X + 2X^2 - Y.$$

[3] The version presented in [9] does differ slightly from the estimator considered here, because the former uses Dumitrescu's symmetrical sample pairs definition. It leads to a slightly more complicated formula for $\hat{p}$, but the difference in performance is negligible.

[4] This model was first proposed in passing in [7], as part of a method for quantifying the accuracy of cover symmetries.

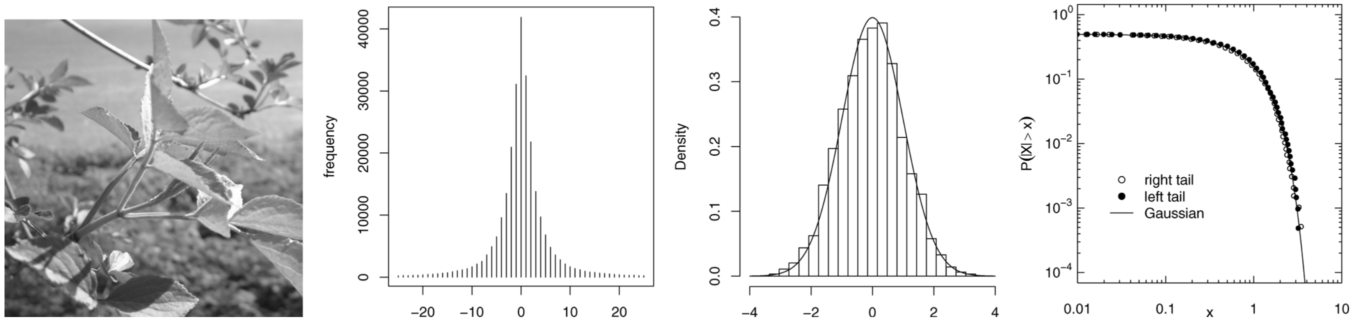

Fig. 4. Experiment with synthetic data. From left to right: the image used; the difference histogram of this image; histogram of standardized $\hat{p}$ computed using synthetically generated $e_m$ and $o_m$, with a standard Gaussian density superimposed; logarithmic tail plot compared with Gaussian tail.

It is already known that $X$ has a Gaussian distribution, but the other terms do not. Therefore, the second-order approximation to $\hat{p}$ is not Gaussian. Rather than proceed to a complex derivation of the exact distribution of $\hat{p}$, simply observe that the contribution to distributional shape by $X^2$ and $Y$ is small—their variance, and covariance, are all $O(N^{-2})$—whereas their contribution to location is $O(N^{-1})$. Therefore, a simple approximation is to ignore all except the shift in mean caused by the additional term $2X^2-Y$. Using $E[Z_m^2]=1, E[X^2]=v(\boldsymbol{d})$ and $E[Y] = 4\sum_m d_{2m+1}/\sum_m(d_{2m+2}-d_{2m})^2$ can be derived. The second approximation to the distribution of $\hat{p}$ is therefore approximate Gaussian

$$\hat{p} \approx \mathrm{N}\big(\mu_2(\boldsymbol{d}), v(\boldsymbol{d})\big), \quad \mu_2(\boldsymbol{d}) = 2v(\boldsymbol{d}) - \frac{4\sum\limits_m d_{2m+1}}{\sum\limits_m (d_{2m+2}-d_{2m})^2} \tag{10}$$

and $v(\boldsymbol{d})$ is as in (9). The results of Section V will bear out the approximations that have been made here, and the necessity of the more complex second approximation.

## V. Experimental Results

These results are tested empirically, computing the LSM estimates over a large set of cover images. The primary test set of covers is 3000 never-compressed bitmaps, downloaded from http://photogallery.nrcs.usda.gov; originally very high resolution color images, for most of the testing they were reduced to approximately 640 × 450 pixels. Tests were repeated using images reduced to grayscale, and also extracting the color channels and using them separately. Additionally, a summary of results for wider tests, with other sets of covers, will be reported.

To test the theory, standard statistical tests of mean and variance will be used. In order to test whether data fit a Gaussian distribution, the Anderson-Darling test [15] will be used; this is known to be generally powerful with particular discriminating power in the tails of the distribution. The tails are especially important if high reliability is the aim, and these will be augmented by plots of both the empirical histogram (which effectively checks the center of the distribution) and a logarithmic plot of the observed distribution function (which exposes any heavy-tailed behavior), compared with the standard Gaussian.

### A. Results From Synthetic Data

First, some synthetic simulations will be reported, testing the accuracy of the results of Section IV-B independent of the accuracy of the cover model in Section IV-A.

Consider initially the first-order approximation (9). Taking a single grayscale image (pictured in Fig. 4), the difference histogram (also displayed) was extracted. For this particular vector $\boldsymbol{d}$, (9) predicts $v(\boldsymbol{d}) = 8.941 \times 10^{-5}$. Two-thousand simulations were then repeated, setting each $e_m$ according to a binomial random variable with parameters $d_m$ and 1/2, and $o_m = d_m - e_m$, then computing $\hat{p}$ according to the LSM algorithm. Standardizing, the theory predicts a Gaussian distribution with zero mean and unit variance for $\hat{p}/\sqrt{v(\boldsymbol{d})}$. This histogram, and logarithmic tail plot, is displayed in Fig. 4.

Close accordance with the theory is seen. The data easily pass the Anderson–Darling normality test ($p = 0.637$). The observed mean is $-1.634 \times 10^{-4}$, not significantly different from zero ($t$-test $p = 0.433$). The observed variance is 8.693 $\times 10^{-5}$, not significantly different from the theoretical prediction of 8.941 $\times 10^{-5}$ ($\chi^2$-test $p = 0.366$). Observe that the tail of the standardized estimates, displayed in Fig. 4, fits a standard Gaussian tail very closely. The first-order approximation to the distribution of $\hat{p}$ has been quite adequate.

However, the experiment was also repeated by artificially reducing the size of each $d_m$ by a factor of 20, simulating a smaller image with the same general characteristics. Another set of charts is not shown; it suffices to say that the data still pass a normality test ($p = 0.059$), and the variance is not significantly different from the prediction ($p = 0.145$). But the observed mean of $-0.00617$ is significantly lower than zero ($p < 10^{-10}$). However, the second approximation (10) would imply a mean of $-0.00515$, not significantly different from the observed value ($p = 0.270$).

This illustrates that the second approximation is necessary for what would be a smaller image, and that it accords well with the empirical results in this case.

### B. Testing the Cover Model

This section considers genuine cover images. First, the cover image model of Section IV-A is tested in isolation. According to this model, the statistic $z_m = (e_m - o_m)/\sqrt{e_m + o_m}$ should have standard Gaussian distribution. This statistic is computed for a set of 3000 grayscale images and the results are displayed, for $m = 1$ and $m = 5$, in Fig. 5. Observe that the model seems appropriate for $m = 5$ (it passes the normality test with $p = 0.276$) but is completely inappropriate for $m = 1$ ($p < 10^{-10}$).

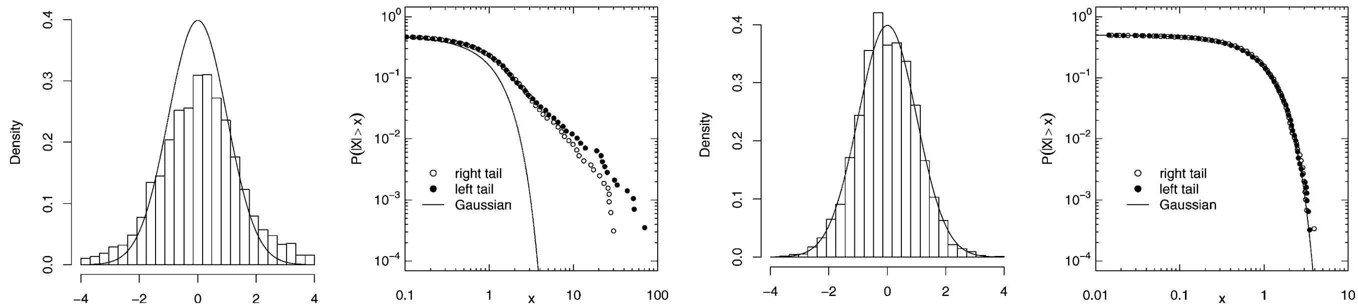

Fig. 5. Tests of the cover image model. From left to right: histogram of $z_1$, with standard Gaussian superimposed; logarithmic tail plot of $z_1$; histogram of $z_5$; logarithmic tail plot of $z_5$.

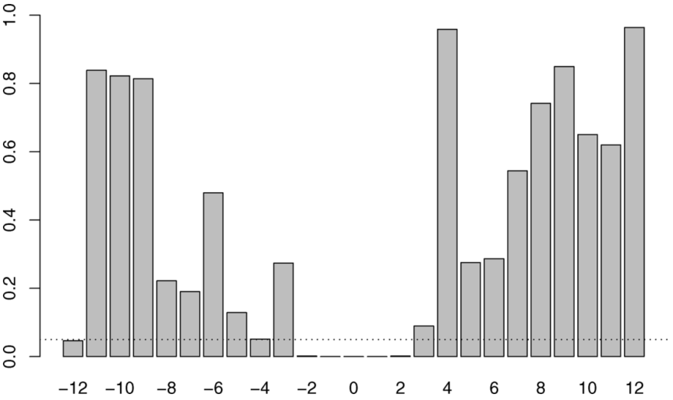

Fig. 6. Tests of the cover image model. Displays the $p$-value for the Anderson–Darling goodness-of-fit test for $z_m$ against a standard Gaussian distribution. $p = 0.05$ is indicated.

Indeed, in the latter case, it would appear to have tails closer to Pareto [14] than Gaussian.

Rather than repeat such charts for every $m$, only the $p$-value for the Anderson–Darling test is displayed, for $|m| \leq 12$, in Fig. 6. It appears that there is no evidence to reject the model for $|m| \geq 3$. Other experiments using covers made of individual color channels extracted from 3000 color images (chart not displayed) bring into question the validity of the model for $|m| = 3$ also. But for $|m| > 3$, the model fits well.

Another experiment was performed using covers which had previously been subject to JPEG compression, in the firm expectation that the frequency-domain quantization would strongly disrupt any assumptions about parity structure. Suprisingly, in fact, it was found that the model still fits for $|m| > 3$ if the JPEG covers are converted to grayscale before use: an unlooked-for bonus. The model is not appropriate for single color channels extracted from previously JPEG-compressed images although, even here, it was found that there were circumstances highly dependent on the nature of the image before compression, in which the fit was reasonable. This is worthy of further study but, for now, the model will only be applied to never-compressed images.

Note that the cover image model is only used for odd values of $m$: the model for $e_{2m+1} - o_{2m+1}$ drives component $m$ of (7). Because the model has been observed to fail for $z_m = -3, -1, 1, 3$, the LSM estimator will henceforth be modified to exclude the corresponding components ($m = -2, -1, 0, 1$). The estimator should then satisfy (9) or (10).

The assumption that the random variables $z_m$ are independent must also be tested. Computing pairwise correlation coefficients between each $z_m$, it was found that none were correlated with $R^2$ of more than 0.1, and for most pairs the correlation was almost exactly zero. The only exception is that $z_1$ and $z_{-1}$ are strongly negatively correlated ($R^2 = 0.84$). Some visual inspections of the data were conducted to test for the possibility of nonlinear dependence—none were found. As long as at least one of $m = -1$ and $m = 0$ is excluded from the sum–square error, it appears that independence of the components may be assumed.

### C. Distribution of the LSM/SPA Estimator

Predictions of between-image error distribution can now be tested for the modified estimator. According to (9), computing $\hat{p}$ and $\boldsymbol{d}$ for each image, the theory predicts that $\hat{p}/\sqrt{v(\boldsymbol{d})}$ is standard Gaussian. However, when the set of 3000 grayscale images was tested, it was found that this was not a very good fit. The histograms are not displayed; it is sufficient to note that the observed mean of this standardized estimate was $-0.326$ (significantly different from the prediction of 0, $p < 10^{-10}$); also, the standardized distribution fails a normality test ($p < 10^{-10}$). It appears that the first-order approximation is not sufficient. Although the images are the same size as the first synthetic experiment in Section V-A, by excluding $m = -2, -1, 0, 1$, many of the pixels are ignored (an average of about 50% of the pixels in each image are now excluded; in some images as many as 90%) and, thus, the available evidence is analogous to that of smaller images.

In fact, the first-order approximation is sufficient for larger images: another set of 3000 images, sized approximately 1.5 M pixels each, gave results passing all of the tests. However, for the images sized 640 × 450, the second approximation (10) will be applied. The theory predicts that $\big(\hat{p} - \mu_2(\boldsymbol{d})\big)/\sqrt{v(\boldsymbol{d})}$ has a standard Gaussian distribution: the histogram and logarithmic tail-plot of these standardized estimates appears in Fig. 7. The standardized mean is 0.0199 (not significantly different from 0, $p = 0.285$), and the standardized variance is 1.049 ($p = 0.0597$). The data pass the Anderson–Darling test with $p = 0.314$. Thus, the second approximation (10) accords very well with the experimental data. Although there appears to be one outlier in the right tail, the reader is cautioned against placing too much significance on the last few data points in a logarithmic tail plot: extreme order statistics are notoriously unreliable measurements.

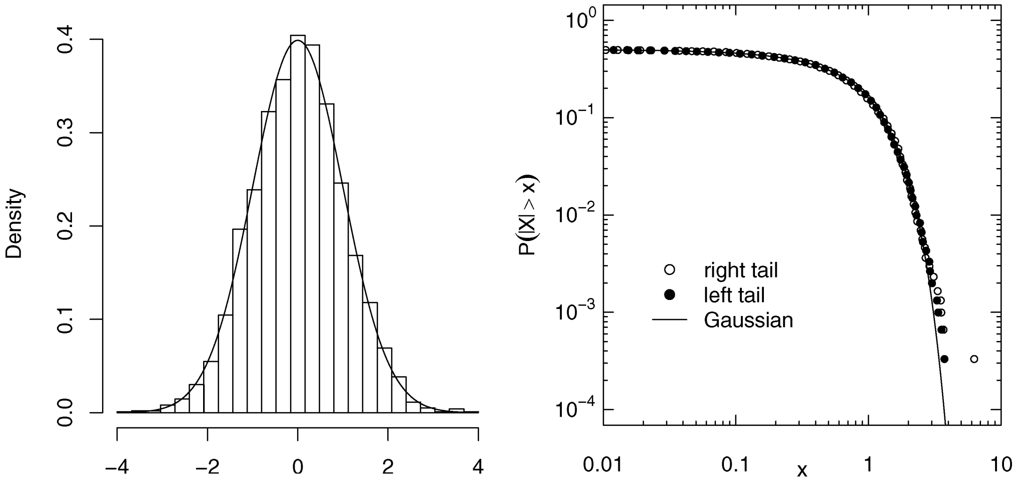

Fig. 7. Observed modified LSM/SPA estimates (in which components $m = -2, -1, 0, 1$ are excluded) in 3000 grayscale covers, standardized according to (10). Left, histogram; Right, logarithmic tail plot. An excellent fit is observed (the single outlier should be disregarded).

This experiment has been repeated for a number of different sets of covers. Only a summary of the results is reported. When single color channels are used (simulating LSB replacement in color images), there is still a good fit for a Gaussian distribution and standardized variance of 1, but it was observed that, on occasion, the mean of the standardized estimates is a little away from zero. This "bias" is less than 0.1, which is statistically significant but not very substantial given that the data are otherwise standard Gaussian. Similar results arise in a set of 3000 smaller images (320 × 225), and with a set of 1000 large raw images converted to grayscale directly from a variety of digital cameras (with no resizing). The slightly inaccurate mean would not greatly damage the calculation of a $p$-value by a steganalyst. In particular, there are no extreme outliers which, in the non-standardized estimates, would cause unavoidable false positive results. Similar results are again obtained when a set of 10000 previously JPEG-compressed, grayscale covers were used (almost regardless of the quality factor used in compression), but extracting single color channels from JPEG-compressed covers gave rise to non-Gaussian results. Clearly, this is due to the failure of the cover model.

The same experiments were carried out without removing $m = -2, -1, 0, 1$ but a Gaussian fit was observed: it has already been demonstrated that the model for cover images does not hold well here, and this merely verifies that the failure carries through into the distribution of the estimator. Removing these components does have a negative impact on the estimator, increasing its variance by not making full use of the data in the stego image. But what is lost in general accuracy may be offset by the removal of outliers using this new theory. Detailed benchmarking of modified detectors to further work is postponed, but initial results suggest that use of the $p$-value allows for much lower false positive rates than previously.

Finally, consider Fig. 1 in the introduction of this paper; for the results to hold exactly, one must change to the modified LSM estimator which excludes $m = -2, -1, 0, 1$, but the histogram of the modified estimator (not displayed) has very much the same shape. Now it is known that an (approximate) Gaussian mixture is being observed. For the same set of 3000 grayscale covers, histograms of the mixture parameters $\mu_2(\boldsymbol{d})$ and $v(\boldsymbol{d})$, with a scatterplot showing how they are correlated, are shown in Fig. 8. The image-specific bias $\mu_2(\boldsymbol{d})$ is almost always negative (the largest observed value was 0.00011 and 99.3% are negative), explaining the negative bias in Fig. 1. The variance $v(\boldsymbol{d})$ is long tailed: some images have very high variance. This, along with some outliers in the bias, accounts for the long tails in Fig. 1. Although the long tails in $v(\boldsymbol{d})$ and $\mu_2(\boldsymbol{d})$ are themselves an as-yet unexplained phenomenon, it is important to note that the quantities $v(\boldsymbol{d})$ and $\mu_2(\boldsymbol{d})$ can be observed from the cover image; those images with a propensity for large steganalysis error can now be identified.

## VI. Applications

An immediate application of a sound model for estimator error is a measure of confidence for the steganalyst. It would be desirable to provide not only an estimate but a confidence interval for the size of hidden payload. However, the theory does not enable going that far, for two reasons. First, only the between-image error (i.e., the theory only applies directly to cover images) has been identified: while the within-image error is negligibly small, relative to between-image error, for small payloads, this is not so for large embedding rates [6]. Second, computing the expected bias and variance of the between-image error requires knowledge of $\boldsymbol{d}$, which is a property of the cover. Of course, the steganalyst does not have access to both cover and stego object.

However the theory is sufficient to form the most important measure of confidence: the $p$-value of the hypothesis test that no payload is present versus some payload is present. The $p$-value is the (im)probability of the observation, given the null hypothesis (i.e., assuming that the object under consideration is itself a cover). Therefore, the observed $\boldsymbol{d}$ can be used to compute $\left(\hat{p} - \mu_2(\boldsymbol{d})\right)/\sqrt{v(\boldsymbol{d})}$ as a standard statistic, knowing that this is standard Gaussian under the null hypothesis; one should be prepared for a small bias up to 0.1 and take care to use the method only on "well-behaved" images (either never-compressed, or previously JPEG-compressed grayscale). And the steganalyst must use the modified LSM/SPA detector which disregards the

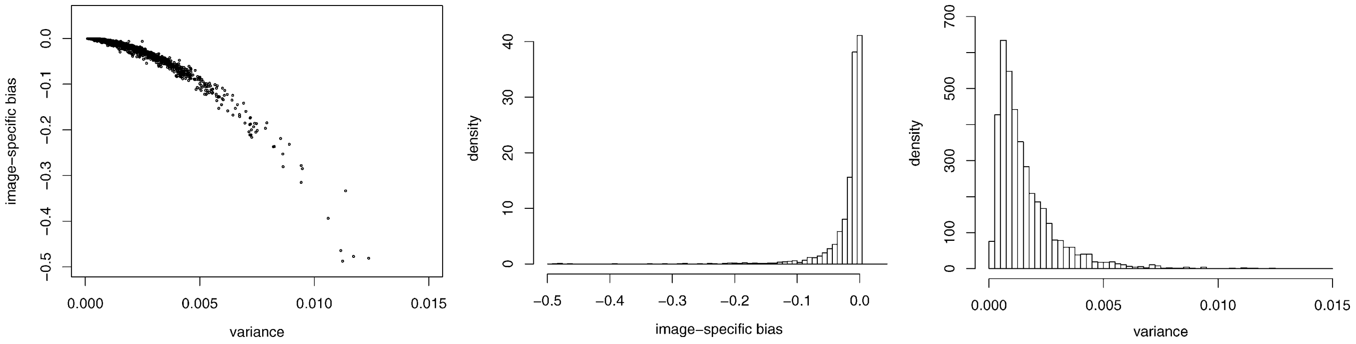

Fig. 8. Observed mixture parameters: image-specific bias $\mu_2(\boldsymbol{d})$ and variance $v(\boldsymbol{d})$ predicted for $\hat{p}$ in 3000 grayscale bitmap images.

components $m = -2, -1, 0, 1$ else the cover model on which the theory is founded cannot be relied upon.

The removal of outliers is seen as the most important contribution of this work: outliers to the right correspond to false positive results and, until now, a certain false positive rate has been almost unavoidable because of the presence of a few stubborn images with a huge positive bias. Computing a true $p$-value removes this problem and paves the way for genuinely high-reliability steganalysis. But it is not yet clear whether what is gained in terms of reduced outliers makes up for the loss of the components $m = -2, -1, 0, 1$, which is effectively to ignore parts of the image; perhaps with a better cover model, such removal will be unnecessary. In any case, the development and benchmarking of improved estimators is postponed to further work. It is reasonable to assume that removal of outliers is a necessity for applications where very low false positive rates are essential.

A second application is in the development of better steganalysis estimators. Consider a weighted least squares method, in which a weighted sum

$$\sum w_m \Big( \frac{s'_m + pt'_m + p^2 u'_m}{(1-p)^2} \Big)^2$$

[cf. (7)] is minimized to find $\hat{p}$. The theory can be applied to determine the weights vector $\boldsymbol{w}$ which gives rise to an estimator with the lowest between-image variance.

Detailed discussion of this is postponed to further work, but elementary calculations show that the optimal weight is given by $w_m = 1/d_{2m+1}$. In practice, this achieves a reduction of around 20% in between-image variance, but at the cost of increased bias (which can now be corrected for).

A more speculative application is to aid the steganographer in selection of a cover image. In the (probably unrealistic) scenario where a steganographer needs to make but a single covert communication, they could choose a cover which makes steganalysis difficult by picking one with high $v(\boldsymbol{d})$. However, if they can send multiple covers, they should instead split the payload across all covers (see [16]) and it would be dangerous to send only images with high $v(\boldsymbol{d})$, which is itself a signature of sorts. And a discerning steganographer can probably do better by using just about any form of embedding other than LSB replacement.

Finally, note that the variance of the between-image error has been shown to be $O(N^{-1})$ if the shape of the difference histogram is fixed, so that the "secure capacity" of a cover, measured in terms of the steganalyst's ability to discriminate stego objects from cover objects, increases as $\sqrt{N}$. This accords with some of the author's other work [16], which conjectures that steganography capacity, in general, is proportional only to the square root of the total cover size.

## VII. Conclusion

A theoretical model of steganalysis error is valuable, not just for the insights it gives into the robustness of the estimator, and the mathematical roots of any weakness. Apart from explaining the long tails and negative bias in the LSM estimator, some possible applications are noted, even in this analysis which only considers between-image error.

It is believed that this is the first rigorous derivation of its kind, and perhaps sets a template for the derivation of error distributions of other quantitative estimators. The two key components are a model for covers which quantifies deviations from the ideal model driving the steganalysis, and some algebra of probabilities to turn this into a distribution for $\hat{p}$. Some other estimators (e.g., the triples analysis of [4]) should only require an extension of the work in this paper. For detectors not based on LSM, some new algebra of probabilities will be needed.

The most immediate direction for further work is to consider within-image error for this estimator. It is likely that the results of Section II will apply again. The within-image errors are due to (4), and the true distribution of $E'_m$ and $O'_m$ is a small multinomial mixture. The multivariate Gaussian approximation exposes the first obstacle: the components are not independent. This seems to complicate the analysis.

Also, to be addressed is how to estimate $v(\boldsymbol{d})$ and $\mu_2(\boldsymbol{d})$ for situations when the cover is not known to the steganalyst. There seems to be an obvious solution: $\boldsymbol{d}$ can be estimated, using the inverse to (3), observations of the stego object, and the estimate of $p$. But errors in the estimate of $p$ will feed back into errors in estimates of $\boldsymbol{d}$, but the aim was to use the latter to correct the bias of the former. However, it may be possible to break this circularity.

Finally, to tidy up this particular work, the cover model of Section IV-A could be refined so that it works also for $|m| \leq 3$. At first sight, it appears that the lack of independence between the parity of nearby pixels must be accounted for. If a good model for the cover which fits the case of small $m$ could be developed, even if it is not Gaussian, it could, in principle, be included in calculations to determine the resulting distribution of $\hat{p}$, although the algebra might be complex. But if it were to turn out that $e_{2m+1} - o_{2m+1}$ is not Gaussian (which some experimental evidence indicates is probably the case), then the principle of least squares estimation is suboptimal. Moving gradually toward genuine maximum-likelihood estimation of $\hat{p}$ should be viewed as the long-term goal of this research.

## REFERENCES

[1] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and grayscale images," *IEEE Multimedia*, vol. 8, no. 4, pp. 22–28, Oct. 2001.

[2] J. Fridrich, M. Goljan, and D. Soukal, "Higher-order statistical steganalysis of palette images," in *Security and Watermarking of Multimedia Contents V*, E. J. Delp, III and P. W. Wong, Eds., 2003, vol. 5020, ser. Proc. SPIE, pp. 178–190.

[3] S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1995–2007, Jul. 2003.

[4] A. Ker, "A general framework for the structural steganalysis of LSB replacement," in *Proc. 7th Information Hiding Workshop*, 2005, vol. 3727, Lecture Notes Computer Science, pp. 296–311.

[5] A. Ker, "Quantitative evaluation of pairs and RS steganalysis," in *Security, Steganography, and Watermarking of Multimedia Contents VI*, E. J. Delp, III and P. W. Wong, Eds., 2004, vol. 5306, ser. Proc. SPIE, pp. 83–97.

[6] R. Böhme and A. Ker, "A two-factor error model for quantitative steganalysis," in *Security, Steganography and Watermarking of Multimedia Contents VIII*, E. J. Delp, III and P. W. Wong, Eds., 2006, vol. 6072, ser. Proc. SPIE, pp. 59–74.

[7] A. Ker, "Fourth-order structural steganalysis and analysis of cover assumptions," in *Security, Steganography and Watermarking of Multimedia Contents VIII*, E. J. Delp, III and P. W. Wong, Eds., 2006, vol. 6072, ser. Proc. SPIE, pp. 25–38.

[8] M. Goljan, J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications," in *Security, Steganography and Watermarking of Multimedia Contents VIII*, E. J. Delp, III and P. W. Wong, Eds., 2006, vol. 6072, ser. Proc. SPIE.

[9] P. Lu, X. Luo, Q. Tang, and L. Shen, "An improved sample pairs method for detection of LSB embedding," in *Proc. 6th Information Hiding Workshop*, 2004, vol. 3200, Lecture Notes Computer Science, pp. 116–127.

[10] J. Fridrich and M. Goljan, "On estimation of secret message length in LSB steganography in spatial domain," in *Security, Steganography, and Watermarking of Multimedia Contents VI*, E. J. Delp, III and P. W. Wong, Eds., 2004, vol. 5306, ser. Proc. SPIE, pp. 23–34.

[11] B. Roue, P. Bas, and J.-M. Chassery, "Improving LSB steganalysis using marginal and joint probabilistic distributions," in *Proc. ACM Workshop on Multimedia and Security*, 2004, pp. 75–80.

[12] R. Böhme, "Assessment of steganalytic methods using multiple regression models," in *Proc. 7th Information Hiding Workshop*, 2005, vol. 3727, Lecture Notes Computer Science, pp. 278–295.

[13] J. Fridrich and D. Soukal, "Matrix embedding for large payloads," in *Proc. SPIE Security, Steganography and Watermarking of Multimedia Contents VIII*, E. J. Delp, III and P. W. Wong, Eds., 2006, vol. 6072.

[14] J. A. Rice*, Mathematical Statistics and Data Analysis*, 2nd ed. Pacific Grove, CA: Duxbury, 1995.

[15] M. Stephens, "Tests based on EDF statistics," in *Goodness-of-Fit Techniques*, R. D'Agostino and M. Stephens, Eds. New York: Marcel Dekker, 1986.

[16] A. Ker, "Batch steganography and pooled steganalysis," presented at the 8th Information Hiding Workshop, 2006.

**Andrew D. Ker** (M'06) was born in Birmingham, U.K., in 1976. He received the B.A. degree in mathematics and computer science and the D.Phil. degree in computer science from Oxford University, Oxford, U.K., in 1997 and 2001, respectively.

Previously, he was a Junior Research Fellow with University College, Oxford. Currently, he is a Royal Society University Research Fellow with the Computing Laboratory, Oxford University. His initial work was in the foundations of computer science and his research interests are steganography and steganalysis.

Dr. Ker is a member of the SPIE, The International Society for Optical Engineering.