Programming Research Group

DERIVATION OF ERROR DISTRIBUTION IN LEAST SQUARES STEGANALYSIS

Andrew D. Ker

PRG-RR-06-03

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD

Abstract

We consider the *Least Squares Method* (LSM) for estimation of length of payload embedded by Least Significant Bit (LSB) 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 work we derive (approximations for) the estimator distribution over cover images: 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, 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. To our knowledge this is the first derivation of steganalysis estimator performance.

1 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 e.g. [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) we probably cannot assume that the steganalyst has much information of this sort.

In this paper we consider a particular quantitative detector for LSB replacement in grayscale images: the *Least Squares Method* (LSM) variant [9] of *Sample Pairs Analysis* (SPA) [3]. Our aim is to *derive* its error distribution; we will be able to do so for one source of error, as long as the detector is modified to remove dependence on a pathological component.

In [6] we showed that steganalysis estimator error should be decomposed into two components: *within-image error* and *between-image error*. In [6] these are separated and their nature investigated empirically for a number of LSB replacement estimators,

Manuscript version 28 July 2006.

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 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, we focus only on between-image error, which is the error in the estimator when there is no payload embedded[1]. Our 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 has some theory which makes examination of steganalysis error. However it does so in passing (as part of the tuning process for the estimator), only for within-image error, and it is not clear that the theory has any connection with experimental practice. To our knowledge we present here the first derivation of steganalysis error which does not make unrealistic assumptions about the source of cover objects, and which accords well with experimental results. This work has applications both in improved steganalysis and, more speculatively, in adaptive steganography.

As an introductory example, we display in Fig. 1 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). We highlight two features 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. Our theory will explain these features in full: 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 the paper is as follows. In Sect. 2 we do some simple mathematics relating to perturbations in parametric curves of a certain type, which will be a key part of the later derivations. In Sect. 3 we describe the LSM/SPA method in just enough detail for our purpose of deriving its error when no payload is embedded. In Sect. 4 we propose a simple model for cover images which explains the errors, and combine this with the previous results to derive first- and second-order approximations to the between-image error distribution. We verify that the second approximation gives a high degree of accuracy in Sect. 5. Briefly we look at applications of this work, in Sect. 6, including a modification of the LSM/SPA method with improved performance. Finally, we draw conclusions in Sect. 7.

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

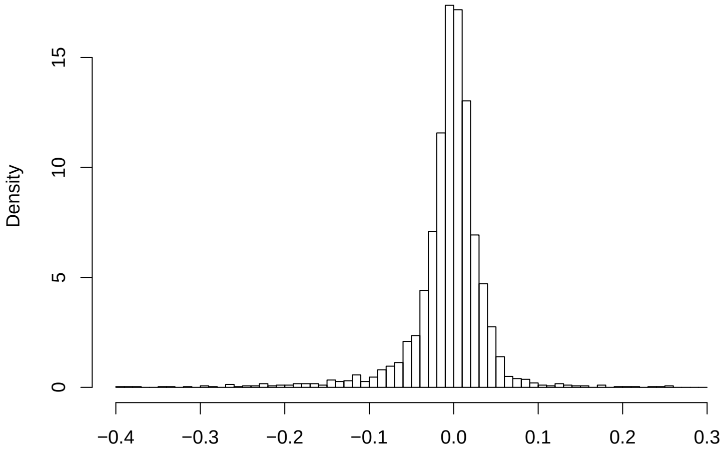

Figure 1: Histogram of observed LSM/SPA estimates of proportionate length of hidden payload, when none is hidden in 3000 grayscale cover images.

2 Small Perturbations in Quadratic Paths

We begin with some abstract mathematics. We will call a parametric curve in $\mathbb{R}^m$ a *quadratic path* if each co-ordinate 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). We will write its locus in the form

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

for $p < 1$, where the vectors $\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{u}$ are in $\mathbb{R}^m$. We are interested in 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. In our application we will want to estimate $\hat{p}$, the value of the parameter on the perturbed curve which is closest to the origin. We will find first- and second-order approximations for $\hat{p}$.

Suppose that a quadratic path P passes through the origin at $p = 0$, and that a perturbed path is P'. Let us 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'. We will approximate P' close to $p = 0$ by its tangent at $p = 0$, which passes through $\boldsymbol{s'}$ and has direction vector $\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}p}|_{p=0} = \boldsymbol{t'} + 2\boldsymbol{s'}$. This is closest to the origin at the point whose vector is orthogonal to the direction vector of the tangent (see Fig. 2), so $(\boldsymbol{s'} + \hat{p}(\boldsymbol{t'} + 2\boldsymbol{s'})).(\boldsymbol{t'} + 2\boldsymbol{s'}) = 0$, which occurs when

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

Now let us identify the perturbations $\boldsymbol{s'} = \boldsymbol{\delta_s}$, $\boldsymbol{t'} = \boldsymbol{t} + \boldsymbol{\delta_t}$, $\boldsymbol{u'} = \boldsymbol{u} + \boldsymbol{\delta_u}$. We have

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

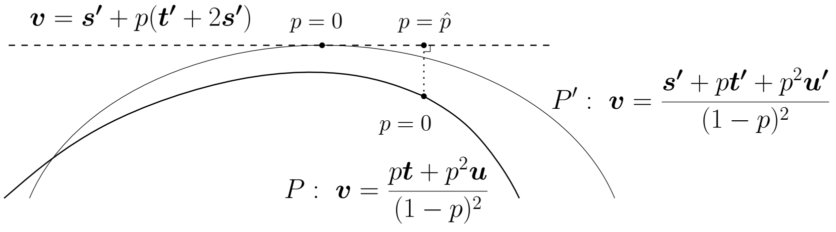
$\boldsymbol{v} = \boldsymbol{s'} + p(\boldsymbol{t'} + 2\boldsymbol{s'})$
$p = 0$
$p = \hat{p}$
$P' :\ \boldsymbol{v} = \dfrac{\boldsymbol{s'} + p\boldsymbol{t'} + p^2\boldsymbol{u'}}{(1-p)^2}$
$p = 0$
$P :\ \boldsymbol{v} = \dfrac{p\boldsymbol{t} + p^2\boldsymbol{u}}{(1-p)^2}$

Figure 2: Small perturbations in quadratic paths.

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

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

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, we obtain

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

These results will be applied in Sect. 4.

3 The Least Squares Method for Steganalysis of LSB Replacement

The Least Squares Method [9] is a quantitative detector for LSB replacement steganography. 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. Its fits into the *structural framework* of [4]: there is a macroscopic property of stego images which depends on the proportionate (as a fraction of the capacity) amount of hidden payload p, a vector $\mathbf{S}(p)$; determination of how $\mathbf{S}(p)$ depends on p and $\mathbf{S}(0)$ and inversion to see how $\mathbf{S}(0)$ depends on $\mathbf{S}(p)$ and p; finally, a model for cover images, expressed in terms of $\mathbf{S}(0)$. Observing $\mathbf{S}(p)$, the estimator for p is the value which implies $\mathbf{S}(0)$ closest to the cover model.

We will describe this estimator in the compact presentation suggested by [4], including only enough detail for our subsequent analyses. As in [4] we will use calligraphic letters ($\mathcal{X}$) for sets, upper-case letters (X) for random variables, and lower-case letters (x) for constants and realisations 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 samples $s_1, s_2, \ldots, s_N$ taking values in the range $0 \ldots 2M + 1$ (typically $M = 127$). A *sample pair* is a pair (s_i, s_j) for some $1 \leq i \neq j \leq N$. Let $\mathcal{P}$ be a set of sample pairs; we will use the set of all pairs which come from horizontally or vertically adjacent pixels (as

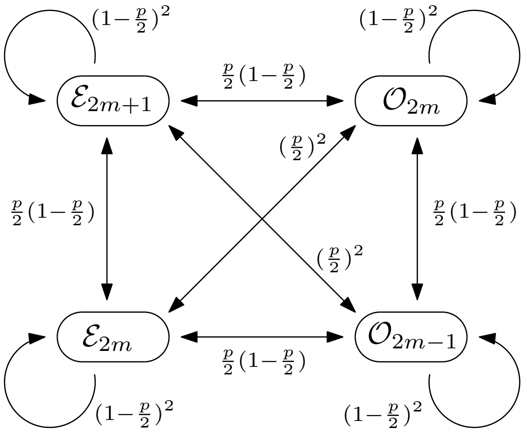

Figure 3: Transitions between trace subsets when proportionate payload p is embedded by LSB replacement.

in [3]). We then consider some subsets of $\mathcal{P}$:

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

in the first of these, $-M \leq m \leq M$; for the second, $-2M \leq m \leq 2M+1$; for the third, $-2M+1 \leq m \leq 2M$.

We suppose that the payload of length pN is a random bitstream (it suffices to be uncorrelated with the cover, so this assumption is not a strong one) embedded using LSB replacement of a random selection of samples, independent of the content of the cover or payload. Then each sample in each pair is altered, independently, with probability $\frac{p}{2}$. The sets $\mathcal{C}_m$ do not involve the least significant bits of the pairs, so any pair in $\mathcal{C}_m$ must remain there after LSB overwriting. The sets $\mathcal{E}_m$ and $\mathcal{O}_m$ we call *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 amongst these four trace subsets according to the transition diagram Fig. 3.

We 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:

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

[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 Least Squares Method of [9]. Their definition is symmetrical in the sample pairs but introduces an unnecessary special case at $m = 0$. This is explained in [4].

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 $p \neq 1$: the inverse is $\frac{1}{(1-p)^2} M\big(1-\frac{p}{2}, -\frac{p}{2}\big)$.

We need two assumptions. First, appealing to the Law of Large Numbers, that the observed realisations 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:

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

Approximate equations of the form of (5) are termed *symmetries* in [7]. They are justified because we expect 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] we will find it 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

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

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 minimises the sum square error of all of these approximately zero quantities. We will not include, here, the mechanics of how such a p may be determined, as this may be found already in [9][3] and is not relevant to our subsequent analysis.

We made two assumptions: (4) and (5). The former is responsible for within-image error, the latter for between-image error. As stated in Sect. 1, in this work we will disregard the within-image error and concentrate 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

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

where

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

[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.

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. We see that $\hat{p}$ is the parameter where the quadratic path $\boldsymbol{v} = \frac{\boldsymbol{s'}+p\boldsymbol{t'}+p^2\boldsymbol{u'}}{(1-p)^2}$ is closest to the origin.

4 Derivation of Between-Image Error

We now derive approximations 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 Sect. 3, but subject to random "error". But we will not change notation, so the reader is warned that some lowercase letters are now random variables.

4.1 A Model for Symmetry Deviation

The assumption $e_m \approx o_m$ is natural, but it does not hold precisely in images. We seek a model for cover images which explains the deviations from exact equality. We would like the model to be as gentle as possible, so that it is not too dependent on the image source for its accuracy.

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 $\frac{1}{2}$, independently of other pairs.

That is, we assume that the *difference histogram* (the frequency of differences of adjacent pixels) is fixed, and that 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. We will make an isolated test of this model in Subsect. 5.2: it will be seen to be quite accurate for $|m| > 3$, marginally so when $|m| = 3$, and not accurate for $|m| \leq 2$. We will do no more than restrict our 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 Generalised Gaussian distribution). However we resist this temptation for now: the more imposing the assumption, the less widely applicable it will be.

Given our model, d_m are constants but the e_m are binomial random variables; e_m then determines o_m. Making the Gaussian approximation $e_m \sim \mathrm{Bi}(d_m, \frac{1}{2}) \approx \mathrm{N}(\frac{d_m}{2}, \frac{d_m}{4})$

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

(valid as long as d_m is at least about 10; see e.g. [11]) we have

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

We only need to use the model for odd 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 the randomness in deviations from the exact equations $e_{2m+1} = o_{2m+1}$.

4.2 Distribution of Between-Image Error

We now return to the geometric version of the LSM estimator. It is the parameter which places $\boldsymbol{v} = \frac{\boldsymbol{s'}+p\boldsymbol{t'}+p^2\boldsymbol{u'}}{(1-p)^2}$ closest to the origin. Let us write $\boldsymbol{s'} = \boldsymbol{s} + \boldsymbol{\delta_s}$, and so on, where $\boldsymbol{s}$, $\boldsymbol{t}$, $\boldsymbol{u}$ 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) we derive

$$\begin{array}{rclcrcl} \boldsymbol{s} & = & \mathbf{0} & \qquad & \boldsymbol{\delta_s} & = & \boldsymbol{\varepsilon} \\ \boldsymbol{t}_m & = & \frac{1}{2}(d_{2m+2} - d_{2m}) & \qquad & \boldsymbol{\delta_t} & = & -\boldsymbol{\varepsilon} \end{array}$$

(we do not need to know $\boldsymbol{u}$ or $\boldsymbol{\delta_u}$). The first-order approximation (1) gives

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

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) simplifies because, here, $\boldsymbol{\delta_t} + 2\boldsymbol{\delta_s} = \boldsymbol{\varepsilon}$. Write $X = \frac{\boldsymbol{\varepsilon}.\boldsymbol{t}}{\boldsymbol{t}.\boldsymbol{t}}$ and $Y = \frac{\boldsymbol{\varepsilon}.\boldsymbol{\varepsilon}}{\boldsymbol{t}.\boldsymbol{t}}$. Then (2) reduces to

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

We already know 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 $\hat{p}$, we will simply note 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 we will ignore all except the shift in mean caused by the additional term $2X^2 - Y$. Using $E[Z_m^2] = 1$ we can derive $E[X^2] = v(\boldsymbol{d})$ and $E[Y] = \frac{4\sum_m d_{2m+1}}{\sum_m (d_{2m+2}-d_{2m})^2}$. Our second approximation to the distribution of $\hat{p}$ is therefore *approximate* Gaussian

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

and $v(\boldsymbol{d})$ is as in (9). The results of Sect. 5 will bear out the approximations we have made here, and the necessity of the more complex second approximation.

5 Experimental Results

We test these results empirically, computing the LSM estimates over a large set of cover images. Our primary test set of covers is 3000 never-compressed bitmaps, downloaded from `http://photogallery.nrcs.usda.gov`; originally very high resolution colour images, for most of our testing we reduced them in size to approximately 640 × 450 pixels. We repeated tests using images reduced to grayscale, and also extracting the colour channels and using them separately. Further, we report a summary of results for testing more widely with other sets of covers.

We will often want to know whether data fits a Gaussian distribution. We will use the Anderson-Darling test [12], which is known to be a generally powerful test with particular discriminating power in the tails of the distribution. The tails are especially important if high reliability is the aim, and we will augment these tests by plots of both the empirical histogram (which effectively checks the centre of the distribution) and a logarithmic plot of the observed distribution function (which exposes any heavy-tailed behaviour), compared with the standard Gaussian.

5.1 Results from Synthetic Data

We begin with some synthetic simulations to test the accuracy of the results of Subsect. 4.2 independently of the accuracy of the cover model in Subsect. 4.1.

We begin with the first-order approximation (9). Taking a single grayscale image (pictured in Fig. 4) we extracted the difference histogram (also displayed). For this particular vector $\boldsymbol{d}$, (9) predicts $v(\boldsymbol{d}) = 8.941 \times 10^{-5}$. We then repeated 2000 simulations, setting each e_m according to a binomial random variable with parameters d_m and $\frac{1}{2}$, and $o_m = d_m - e_m$, then computing $\hat{p}$ according to the LSM algorithm. Standardising, we expect to see 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.

We see close accordance with the theory. The data easily pass the Anderson-Darling normality test ($p = 0.637$). The observed mean is -1.634×10^{-4}, not significantly different from zero (t-test $p = 0.433$). The observed variance is 8.693×10^{-5}, not significantly different from the theoretical prediction of 8.941×10^{-5} (χ^2-test $p = 0.366$). In Fig. 4 we see that the tail of the observed standardised estimates follows a standard Gaussian tail very closely. The first-order approximation to the distribution of $\hat{p}$ has been quite adequate.

However we also repeated the experiment by artificially reducing the size of each d_m by a factor of 20, simulating a smaller image with the same general characteristics. We do not show another set of charts, but comment that 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 ($p < 10^{-10}$). We must go to the second approximation (10), which would imply a mean of -0.00515, not significantly different from the observed value ($p = 0.270$).

We see 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.

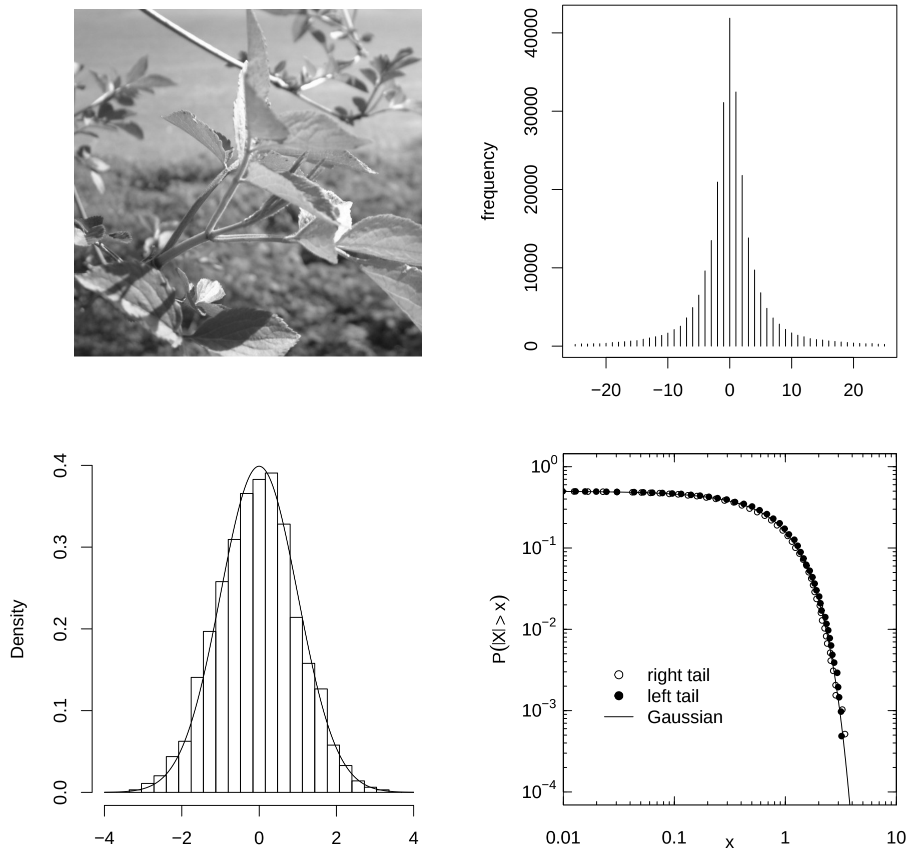

Figure 4: Experiment with synthetic data. Top left, the image used; Top right, the difference histogram of this image; Bottom left, histogram of standardised $\hat{p}$ computed using synthetically-generated e_m and o_m, with a standard Gaussian density superimposed; Bottom right, logarithmic tail plot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

5.2 Testing the Cover Model

We now turn to genuine cover images. First, we test the cover image model of Subsect. 4.1 in isolation. According to this model, the statistic $z_m = (e_m - o_m)/\sqrt{e_m + o_m}$ should have standard Gaussian distribution. We computed this statistic for a set of 3000 grayscale images and display the results, for $m = 1$ and $m = 5$, in Fig. 5. We see that the model seems appropriate for $m = 5$ (it passes the normality test with $p = 0.276$) but completely inappropriate for $m = 1$ ($p < 10^{-10}$). Indeed, in the latter case it would appear to have tails closer to Pareto [11] than Gaussian.

Rather than repeat such charts for every m, we just display the p-value for the Anderson-Darling normality test, 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 colour channels extracted from 3000 colour images (chart not displayed) leads

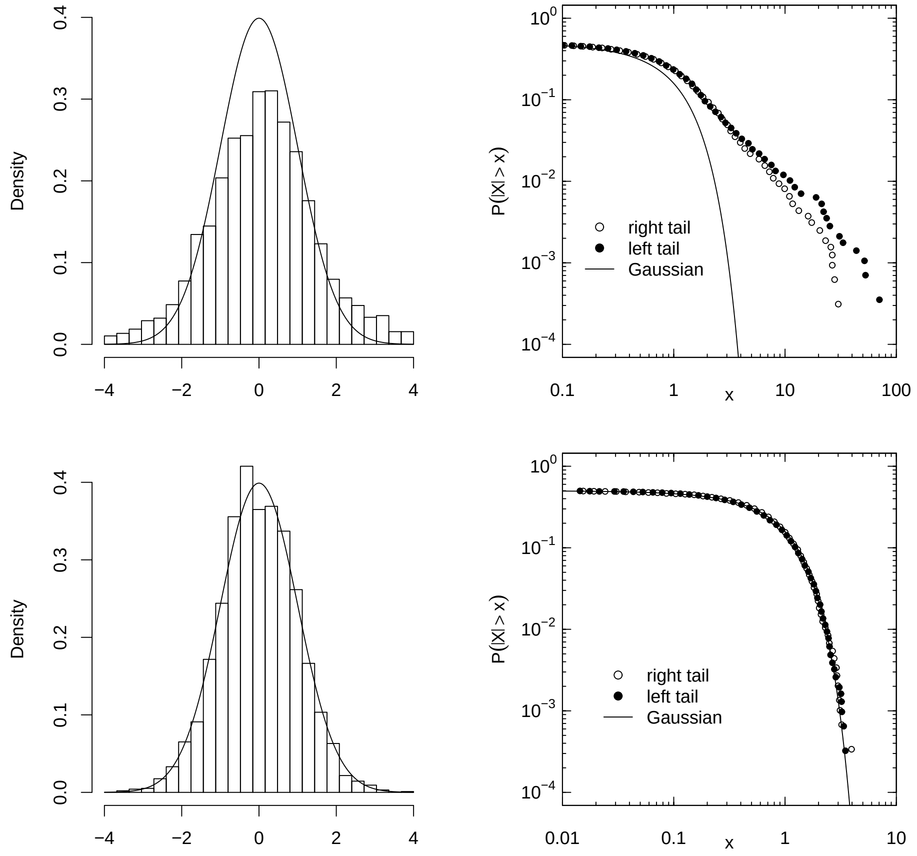

Figure 5: Tests of the cover image model. Top left, histogram of z_1, with standard Gaussian superimposed; Top right logarithmic tail plot of z_1; Bottom left, histogram of z_5; Bottom right, logarithmic tail plot of z_5.

us to 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. In fact, to our surprise, we 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 colour channels extracted from previously JPEG-compressed images, although even here we 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 we apply the model only for its intended use on never-compressed images.

Note that we only *use* the cover image model for odd values of m: the model for $e_{2m+1} - o_{2m+1}$ drives component m of (7). We propose modifying the LSM estimator to *exclude* the components $m = -2, -1, 0, 1$ from the sum-square error computation. The estimator should then satisfy (9) or (10).

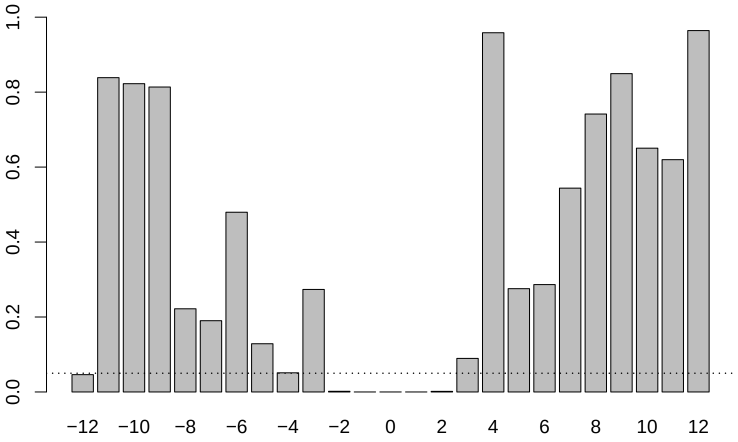

Figure 6: Tests of the cover image model. Displays the p-value for the Anderson-Darling goodness-of-fit test for z_m against a standard Gaussian distribution. $p = 0.05$ is indicated.

We must also test the assumption that the random variables z_m are independent. Computing pairwise correlation coefficients between each z_m we observed that all were either not significantly correlated, or weakly correlated with $R^2 < 0.1$. The only exception is that z_1 and z_{-1} are strongly negatively correlated ($R^2 = 0.84$). As long as least one of $m = -1$ and $m = 0$ is excluded from the sum-square error, we may assume independence of the components.

5.3 Distribution of the LSM/SPA Estimator

With this modified estimator, we are now ready to test our predictions of between-image error distribution. According to (9), computing $\hat{p}$ and $\boldsymbol{d}$ for each image, we should observe that $\hat{p}/\sqrt{v(\boldsymbol{d})}$ is standard Gaussian. However when we tested our set of 3000 grayscale images we found that this was *not* a very good fit. We do not display histograms, but merely note that the observed mean of this standardised estimate was -0.326 (significantly different from the prediction of 0, $p < 10^{-10}$); also, the standardised 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 Subsect. 5.1, by excluding $m = -2, -1, 0, 1$ we are ignoring quite a lot of the pixels (most of the adjacent pixels in an image are close in value) and put ourselves into the case analogous to small images.

We note in passing that the first-order approximation *is* sufficient for larger images: another set of 3000 images, sized approximately 1.5M pixels each, gave results passing all the tests. However for the images sized 640×450 we move on to the second approximation (10). We expect to observe 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 standardised estimates appears in Fig. 7. The standardised mean is 0.0199 (not significantly different from 0, $p = 0.285$), and the standardised variance is 1.049 ($p = 0.0597$). The fit passes the Anderson-Darling test

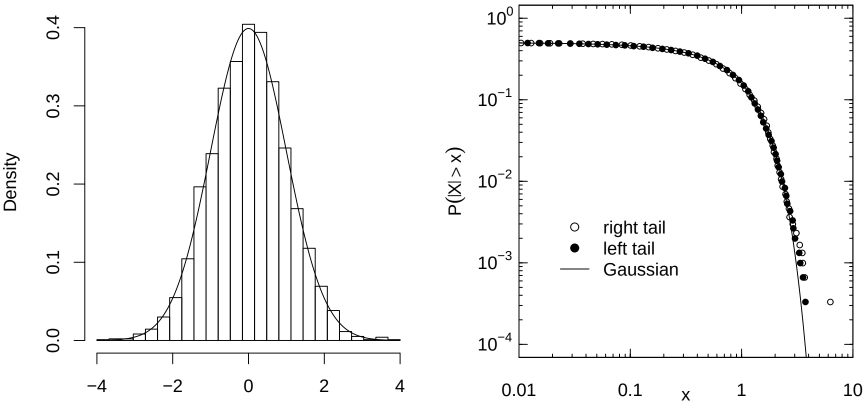

Figure 7: Observed LSM/SPA estimates in 3000 grayscale covers, standardised according to (10). Left, histogram; Right, logarithmic tail plot. An excellent fit is observed (the single outlier should be disregarded).

with $p = 0.314$. The second approximation (10) accords very well with the experimental data. Although there appears to be one outlier in the right tail, we caution the reader against placing too much significance on the last few data points in a logarithmic tail plot: extreme order statistics are notoriously unreliable measurements.

We have repeated this experiment for a number of different sets of covers. We report only a summary of the results. When single colour channels are used (simulating LSB replacement in colour images) we still see a good fit for a Gaussian distribution and standardised variance of 1, but sometimes observe that the mean of standardised 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 the Warden. In particular there are no extreme outliers, which in the non-standardised 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 colour channels from JPEG-compressed covers gave rise to non-Gaussian results. Clearly this is due to failure of the cover model.

The same experiments were carried out *without* removing $m = -2, -1, 0, 1$ but there we observe no Gaussian fit: we already know that the model for cover images does not hold well here, and we have merely verified that it means 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 is gained in allowing us to remove outliers using this new theory. We postpone detailed benchmarking of modified detectors

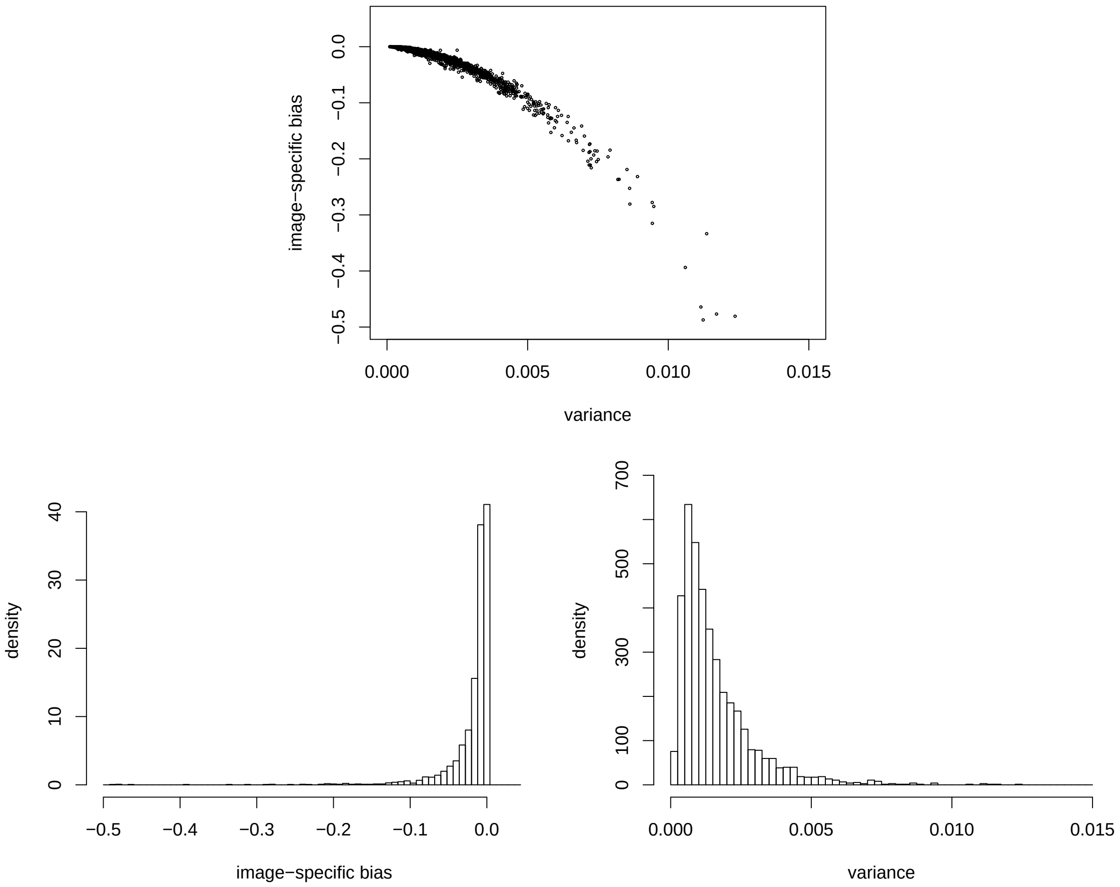

Figure 8: Observed mixture parameters: image-specific bias $\mu_2(\boldsymbol{d})$ and variance $v(\boldsymbol{d})$ predicted for $\hat{p}$ in 3000 grayscale bitmap images.

to further work, but note that use of the p-value allows for much lower false positive rates than previously.

Finally, we return to Fig. 1; we 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. We now know that we are observing an (approximate) Gaussian mixture. For the same set of 3000 grayscale covers we show histograms of the mixture parameters $\mu_2(\boldsymbol{d})$ and $v(\boldsymbol{d})$, and a scatterplot showing how they are correlated, 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.

6 Applications

An immediate application of a sound model for estimator error is a measure of confidence for the steganalyst. We would like to be able to provide not only an estimate but a confidence interval for the size of hidden payload. However our theory does not enable us to go that far, for two reasons. First, we have only identified the between-image error: whilst 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 we can use the observed $\boldsymbol{d}$ for a standardised statistic $\big(\hat{p} - \mu_2(\boldsymbol{d})\big)/\sqrt{v(\boldsymbol{d})}$ and we know that its distribution is standard Gaussian under the null hypothesis, although we should be prepared for a small bias up to 0.1 and must take care to use the method only on "well-behaved" images (either never-compressed, or previously JPEG-compressed grayscale). The steganalyst must use the modified LSM/SPA detector which disregards the components $m = -2, -1, 0, 1$ else the cover model on which the theory is founded cannot be relied upon.

The most important contribution here, we believe, is the removal of outliers. 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.

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

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

(cf. (7)) is minimised to find $\hat{p}$. We can apply the theory to determine the weights vector $\boldsymbol{w}$ which gives rise to an estimator with the lowest between-image variance.

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

A more speculative application is to aid the steganographer in selection of a cover image. There might be many considerations for the steganographer, but one is to choose a cover which makes steganalysis difficult. They can use these results to ensure that the steganalyst must have relatively low confidence in their results, by picking a cover with high $v(\boldsymbol{d})$. (However the steganographer can probably do better by using just about any form of embedding other than LSB replacement.)

Finally, we 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 $\sqrt{N}$. The accords with some of our other work [13], which conjectures that steganography capacity in general is proportional only to the square-root of the total cover size.

7 Conclusions

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 we have noted some possible applications, even of this analysis which only considers between-image error.

We believe that this is the first rigorous derivation of its kind, and perhaps sets a template for 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 Sect. 2 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 we must address the question of how to estimate $v(\boldsymbol{d})$ and $\mu_2(\boldsymbol{d})$, for situations when the cover is not known to the steganalyst. There seems 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 $\boldsymbol{d}$, whereas we would like 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, we aim to refine our cover model of Subsect. 4.1, so that it works also for $|m| \leq 3$. At first sight it appears that we must account for lack of independence between parity of nearby pixels. If a good model for the cover which fits the case of small m could be developed, even if it is not Gaussian, we could in principle include it in our calculations to determine the resulting distribution of $\hat{p}$, although the algebra might be complex. But note that 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. We are gradually moving towards genuine maximum-likelihood estimation of $\hat{p}$, which should be viewed as the long-term goal of this research.

References

[1] J. Fridrich, M. Goljan, and R. Du, “Reliable detection of LSB steganography in color and grayscale images,” ser. Proc. ACM Workshop on Multimedia and Security, 2001, pp. 27–30.

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

[3] S. Dumitrescu, X. Wu, and Z. Wang, “Detection of LSB steganography via sample pair analysis,” in *Proc. 5th Information Hiding Workshop*, ser. Springer LNCS, vol. 2578, 2002, pp. 355–372.

[4] A. Ker, “A general framework for the structural steganalysis of LSB replacement,” in *Proc. 7th Information Hiding Workshop*, ser. Springer LNCS, vol. 3727, 2005, pp. 296–311.

[5] ——, “Quantitive evaluation of Pairs and RS steganalysis,” in *Security, Steganography, and Watermarking of Multimedia Contents VI*, ser. Proc. SPIE, E. J. Delp III and P. W. Wong, Eds., vol. 5306, 2004, 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*, ser. Proc. SPIE, E. J. Delp III and P. W. Wong, Eds., vol. 6072, 2006, pp. 59–74.

[7] A. Ker, “Fourth-order structural steganalysis and analysis of cover assumptions,” in *Security, Steganography and Watermarking of Multimedia Contents VIII*, ser. Proc. SPIE, E. J. Delp III and P. W. Wong, Eds., vol. 6072, 2006, 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*, ser. Proc. SPIE, E. J. Delp III and P. W. Wong, Eds., vol. 6072, 2006, pp. 1–13.

[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*, ser. Springer LNCS, vol. 3200, 2004, 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*, ser. Proc. SPIE, E. J. Delp III and P. W. Wong, Eds., vol. 5306, 2004, pp. 23–34.

[11] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Duxbury Press, 1995.

[12] M. Stephens, “Tests based on EDF statistics,” in *Goodness-of-Fit Techniques*, R. D’Agostino and M. Stephens, Eds. Marcel Dekker, 1986.

[13] A. Ker, “Batch steganography and pooled steganalysis,” to appear in *Proc. 8th Information Hiding Workshop*, 2006.