# STEGANALYSIS OF QUANTIZATION INDEX MODULATION DATA HIDING

*K. Sullivan, Z. Bi, U. Madhow, S. Chandrasekaran, and B.S. Manjunath*

Dept. of Electrical and Computer Engineering
University of California at Santa Barbara
Santa Barbara CA 93106

## ABSTRACT

Quantization index modulation (QIM) techniques have been gaining popularity in the data hiding community because of their robustness and information-theoretic optimality against a large class of attacks. In this paper, we consider detecting the presence of QIM hidden data, which is an important consideration when data hiding is used for covert communication, or steganography. For a given host distribution, we are able to quantify detectability compactly in terms of a parameter related to the robustness of the hiding scheme to attacks. Using detection theory we show that QIM quickly transitions from easily detectable to virtually undetectable as this parameter varies. We also obtain performance benchmarks for QIM hiding in images, indicating that a scheme designed to be robust to, say, a moderate degree of JPEG compression, should be easily detectable. While practical application of detection theory to images is difficult because of statistical variations across images, we employ supervised learning to show that standard QIM schemes for images are indeed quite easily detectable. However, it remains an open issue as to whether it is possible to devise QIM variants that are less vulnerable to steganalysis.

## 1. INTRODUCTION

Quantization Index Modulation (QIM) refers to a class of data hiding schemes that exploit Costa's [1] now famous findings by embedding information in the choice of quantizers. Over the past few years, QIM-based data hiding has received increasing attention from the data hiding community because it is more robust than established techniques such as spread spectrum and least significant bit (LSB) hiding. Recently proposed QIM schemes include Chen and Wornell's QIM and dither modulation [2], Eggers et al's scalar Costa scheme (SCS) [3], and application tailored implementations such as [4, 5, 6, 7].

Given that steganography, or covert communication, is an important application of data hiding, it is natural to ask how easy it is to detect the presence of data hidden using QIM. Thus, the subject of this paper is *steganalysis* (i.e., detection of steganographic communication) of QIM-based hiding. To date, there appears to have been little systematic investigation of this issue, a notable exception being the work of Guillon et al [8] on steganalysis of SCS, based on modeling QIM as inducing additive quantization noise. We employ a more detailed model of QIM in the present work, and apply both detection theory and supervised learning techniques to draw our conclusions.

This research was supported in part by a grant from ONR #N0014-01-1-0380. Email of corresponding author: sullivak@ece.ucsb.edu

Section 2 studies, under idealized conditions, the fundamental limits of steganalysis for QIM. We consider independent and identically distributed (i.i.d.) host (or cover) data, and assume that the steganalyst knows the host distribution. Using hypothesis testing techniques as in [9], we provide performance benchmarks for several variants of QIM. The detectability of QIM for a given host distribution can vary widely, depending on the design level of robustness against attacks. QIM is more easy to detect for distributions of transform domain image coefficients, which exhibit a strong peak at zero which is changed significantly by standard QIM variants. This implies that QIM hiding in images in the transform domain should be easily detectable. In practice, the host distribution for an image is not known, and exhibits significant variations from image to image. However, in Section 3, we show that standard supervised learning techniques using the received distribution as the feature approach detection-theoretic performance limits. In particular, QIM-based hiding designed to resist moderate levels of JPEG compression is quite easily detectable. Lyu and Farid have also used learning systems for steganalysis [10] with promising results. However, since their feature set is chosen without regard to the steganography scheme it is widely applicable, but takes a performance hit due to its generality. A side-by-side comparison would not be illuminating, since they detect non-QIM hiding. Our conclusions are stated in Section 5.

## 2. OPTIMAL DETECTION OF QIM HIDING

The simplest form of quantization based data hiding quantizes the host signal with a quantizer indexed by the message. If $\mathbf{s}$ is the stego signal, $m$ the message, and $\mathbf{x}$ the cover or host signal, we have $\mathbf{s}(\mathbf{x}, m) = q_m(\mathbf{x})$. The stego signal will consist only of values in the set of quantizer outputs. This is appropriate if the signal is expected to be quantized, for compression for example. Dither modulation [2], can produce a stego signal covering all of the values of the host signal. Here the quantizers are shifted according to a changing dither level, i.e. $\mathbf{s}(\mathbf{x}, m) = q_m(\mathbf{x} + \mathbf{d}) - \mathbf{d}$.

There exist more advanced flavors of QIM, which provide advantages to simpler versions. However most practical implementations we have seen use either simple QIM, or dither modulation, with uniform scalar quantizers. We focus on these cases.

Let $P_X(x)$ be the probability mass function (PMF) of the host. We assume $X$ is i.i.d. so the 1-dimensional PMF is sufficient for classification. Since we are using scalar quantizers and i.i.d. data, we will use scalar notation from here on out: $S = s_i, X = x_i$, etc. We can find the PMF of $S$ as a function of $P_X(x)$. We begin with a non-hiding, uniform scalar quantizer. The output levels are the integer multiples of the step-size, $\Delta^*$, and the probability of a given output, $A$, is just the sum of probabilities that are quantized

to that output. Defining the range of input values quantized to a single output value as $\mathcal{X}^*(a) \triangleq [a - \Delta^*/2, a + \Delta^*/2)$ then the PMF is

$$P_A(a) = \begin{cases} P_X(x \in \mathcal{X}^*(a)), a \in k\Delta^* \\ 0 \qquad\qquad \text{else} \end{cases} \tag{1}$$

Where $k$ is any integer. If now a choice of quantizer is used to hide binary data, $B$, we split the original quantizer into 2 coarser subsets, each with step-size $\Delta = 2\Delta^*$. The quantizer associated with sending a 1 is identical to that as for sending 0, but shifted by $\Delta/2$. Assuming the probability of 0 is equal to 1, we have

$$P_S(s) = \begin{cases} \frac{1}{2} P_X(x \in \mathcal{X}(s)) s \in k\Delta/2 \\ 0 \qquad\qquad \text{else} \end{cases} \tag{2}$$

Where $\mathcal{X}(s) \triangleq [s - \Delta/2, s + \Delta/2)$ is the analogous range for the new $\Delta$. Unlike standard quantization, these regions overlap for adjacent values of $s$. We note at this point that if the goal of the steganographer is to mimic an existing quantizer, for example a compression scheme, then the hider can stop here, without using dither modulation. In [4] and [7], the authors use this to imitate the output of JPEG and JPEG2000 respectively. We examine the detection of this first case below.

For dither modulation, we let $D$ be a pseudorandom variable uniformly distributed over $[-\Delta/4, \Delta/4)$ so that the output will cover all the values of the input, and will not leave tell-tale signs of quantization. In this range, $P_D(d) = 2\epsilon/\Delta$ where $\epsilon$ is the granularity of the data. With this dithering, any $s$ is valid, subject to the granularity of the system. For every received $S$ there is one and only one valid value of $d$ that could have made that value of $s$. For any valid $s$, $P_S(s) = P(B = 0, 1) \cap P_X(x \in \mathcal{X}(s)) \cap P_D(d = \text{required})$. Again assuming equiprobable message data and plugging in for $P_D$ we have

$$P_S(s) = \frac{\epsilon}{\Delta} P_X(x \in \mathcal{X}(s)) \tag{3}$$

Armed with equations (1), (2), and (3) we can find the performance of a detector operating in two scenarios. The first is distinguishing between host values that have been quantized versus QIM data embedding (without dithering). The second case is distinguishing between an unquantized host and a host with dithered QIM data embedded.

The optimal detector in the Neyman-Pearson sense of maximizing the probability of detection while maintaining a given false alarm probability is the well known likelihood ratio test [11]:

$$\delta_{L(\mathbf{y})} = \begin{cases} 1 L(\mathbf{y}) \triangleq \left( \frac{P_S(\mathbf{y})}{P_X(\mathbf{y})} \right) \geq \tau \\ 0 \qquad L(\mathbf{y}) < \tau \end{cases}$$

Before we analyze the performance of this detector for some example PMFs, we can gain some insight into what will be detectable simply by inspecting $L(\mathbf{y})$.

**Case I:** Quantized host versus non-dithered QIM hiding

Here we compare to $A$ rather than $X$. The $y_i$ in $\mathbf{y}$ are independent, so $L(\mathbf{y})$ is:

$$L(\mathbf{y}) = \prod_{i=1}^{N} \frac{1/2 \sum_{y_i - \Delta/2 \leq x < y_i + \Delta/2} P_X(x)}{\sum_{y_i - \Delta/4 \leq x < y_i + \Delta/4} P_X(x)}$$

Basically hiding sums over twice the range, and compensates by halving the total. Therefore a smoothly varying PMF will be more difficult to detect than a spikey one.

**Case II:** Non-quantized host versus dither modulation hiding:

$$\prod_{i=1}^{N} \frac{(\epsilon/\Delta) \sum_{y_i - \Delta/2 \leq x < y_i + \Delta/2} P_X(x)}{P_X(y_i)}$$

This is exactly the ratio of the average (over $\Delta$) to the original. Dither modulation hiding therefore acts as a moving average filter on the PMF. Intuitively, host PMFs with high frequency components relative to $\Delta$ will be much easier to detect than a smoothly varying PMF. Indeed, as is noted in [8], a uniformly distributed host would be impossible to detect.

Typically a steganographer will be hiding in data transformed to make it suitable for compression. This data will generally have values concentrated towards the mean. That is, the PMF will tend to have a large spike at the center. See for example the histogram of DCT coefficients of an image in Figure 1. For PMFs such as these, the detectability is strongly linked to the concentration of probablity near the mean compared to the step size of the quantizers, or the ratio of the standard deviation $\sigma$ to $\Delta$. $\Delta$ is directly proportional to the robustness of the hiding.

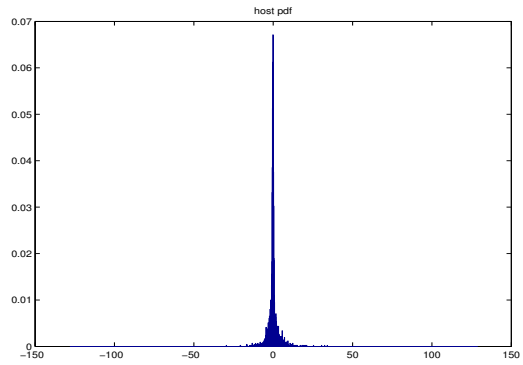


**Fig. 1**. The empirical PMF of the DCT values of an image. The PMF looks not unlike a Laplacian, and has a large spike at zero.

To quantify this observation, we can find the performance of the detector for a given host distribution. We cannot estimate the average probability of error of the detector, because the priors can not be known; who knows how many steganographers exist? As a metric we use the sum of the probabilities of false alarm and missed detection. For a known PMF, we find upper bounds on these probabilities by using Chernoff bounds (for details, see for example [11]). Chernoff bounds allow us to find a bound on the performance even at very low probability of error, which is not possible with simulations. We find the detectability is extremely sensitive to the ratio $\sigma/\Delta$, see Figure 2. Here, we are detecting a Laplacian PMF at rate 1. Within a short range of $\sigma/\Delta$, the detection metric goes from nearly certain detection to almost random detection. Gaussian PMFs have a similar relationship.

The hider then should choose to embed in either a high variance host, or use a small $\Delta$. However the choice of hosts may be limited, and a smaller $\Delta$ will weaken its robustness to external attacks. He or she may choose then to embed less data than is possible in order to avoid detection. We introduce a rate, $R$, measured in bits per host sample to characterize this. For scalar QIM, $0 < R \leq 1$. As $R$ is reduced the detectable difference between the hidden statistics and host statistics is diluted by the host samples that pass unchanged. We can easily adjust equations (2) and (3) to reflect this:

$$P_S(s, R) = R P_S(s, 1) + (1 - R) P_X(s) \tag{4}$$

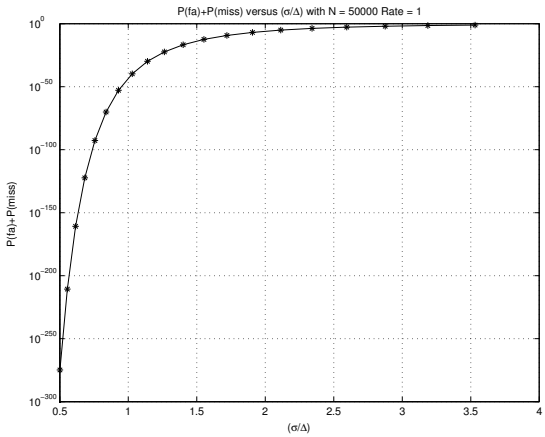


**Fig. 2**. The detector is very sensitive to the width of the PMF.

where $P_S(s, 1)$ is the previous full-embedding stego PMF.

The hypothesis that data is hidden is now composite. To detect this, we use the generalized likelihood ratio test where $L(\mathbf{y})$ is now:

$$L(\mathbf{y}) \triangleq \max_R \left( \frac{P_S(\mathbf{y}, R)}{P_X(\mathbf{y})} \right)$$

To estimate error probabilities with the GLRT, we use computer simulation rather than Chernoff bounds. Hiding at a lower rate certainly decreases the detectability. There is however a catch. The message the sender wants to send covertly has a predetermined length. The lower the rate, the more host samples the hider must use to embed the message. Since this increase in the number of samples increases the steganalyzer's ability to detect the hidden data, the increase in privacy caused by lowering the rate is somewhat offset. Therefore the hider may not be as safe as he or she thought. We illustrate this with an example. Suppose a hider is sending a 15000 bit message in 50000 host samples ($R = .3$) If the host is a Gaussian with $(\sigma/\Delta) = 1$ the detector has an error sum of 0.070. If we hold the number of samples constant but halve the rate to $R = .15$, the sum of errors is 0.366. However this will only send 7500 bits. To send the entire message the hider will have to use 100000 host samples. The performance taking this into account is 0.205.

Finally, in implementing these schemes on real world data, certain adjustments must be made to the basic scheme. For example both [4] and [7] exclude low-valued coefficients from embedding, to avoid visual distortion of the final image. Furthermore, as we mentioned above the host data typically has a characteristically sharp increase of probability near the mean, which will be noticeably smoothed by hiding. Setting a low-valued threshold for embedding also helps avoid this obvious artifact. This thresholding however leaves a new characteristic effect on the PMF near the low-values. The derivation of this modified stego PMF is straightforward but lengthy and is skipped here for brevity.

## 3. STEGANALYSIS WITH SUPERVISED LEARNING

In the steganalytic method described in the previous section, it is assumed that the statistics of the host are known. This is obviously not the case in real world detection. Also, a hiding implementation will often allow a range of step-sizes, $\Delta$, for embedding which is also unknown. The LRT gives us a "best-case" bound on detection. For practical tests, we must assume no prior knowledge.

There is an alternative method, the supervised learning method, which makes no assumptions about the statistics or $\Delta$. Instead, it finds the difference between cover images and images with hidden content entirely from the data through training.

There are usually four steps in any supervised learning procedures: data set construction, feature extraction, training and testing. We describe these steps as follows:

**Data set construction:** For standard supervised learning tasks, a training and a testing set are needed. In our work, we use images from three distinctively different image collections: digital orthophoto quarter-quadrangle (DOQQ) aerial images, Corel PhotoCD (CPCD) images, and images taken with a Canon digital camera (CADC). From each collection, we create a training set and a testing set each consisting of 500 randomly chosen images. Within each set, we hide a random bit stream in half of the images, and therefore, each set contains natural host images and images with hidden content. The task is to distinguish these two classes of images by training a classifier on the training data set and checking the prediction accuracy of the classifier on the testing set.

**Feature extraction:** Before training takes place, raw data, or images in our case, need to be represented by a set of attributes, or features. We assume that our classifier is targeted to a particular implementation of QIM, and choose the features to match. For our testing, we tested on a QIM implementation, [4], that embeds in the 8x8 blockwise DCT coefficents of an image. Since the DCT tends to decorrelate the intensity values, our model of an i.i.d. $X$ as given in the previous section is justified. We therefore use a histogram as an empirical PMF for our feature vector. We compute the histogram with 300 bins over all the coefficients that are typically embedded into, which gives us a 300 dimensional feature vector.

In the QIM implementation we tested on, the hider chooses a $\Delta$ large enough to withstand a pre-determined JPEG compression quantization. We refer to this level as the design quality factor (QF). The smaller the design quality factor, the larger the step-size $\Delta$.

**Training and testing:** We perform the supervised learing in two stages. In the first stage, we train the classifier on a training and testing set from the same image collection. Since images from the same collection are usually similar in content or texture, this is an easier task than the more general case. In the second stage, we create mixed training and testing sets with images from all three collections. This is a more difficult task but more general.

## 4. RESULTS

**With known design quality factor:** In this experiment, we set the design quality factor at 50 to hide data in images in both the training and testing sets, which means when we make a detection, we already know that if there is data hidden in an image, the design quality factor is 50. Both the host images and the images with hidden content are then compressed to JPEG at the same quality factor in order to avoid detection of JPEG compression. The results of detection error for this test are shown in Table 1. We find that if the design quality factor is known, the detection with supervised learning gives very low error rates, which remains low even at severe JPEG compression. But we understand the design quality factor is an extra information which is not usually available for the detector and it is expected to make detection simpler. In the next

| Final QF | 100 | 90 | 80 | 70 | 60 | 50 |
|---|---|---|---|---|---|---|
| DOQQ | 0 | 0 | 0 | 0 | 0 | 0 |
| CPCD | 0 | 0 | .004 | 0 | .044 | .052 |
| CADC | 0 | 0 | 0 | 0 | 0 | .016 |

**Table 1**. If the design quality factor is known (set at 50), a very low detection error can be achieved at all compression rates. Here '0' means no errors occurred in 500 tests so the error rate is $< 0.002$

test, we eliminate this restriction.

**With unknown design quality factor:** We perform this test with an unknown design quality factor. This is achieved by creating a training and testing sets by hiding data in images with the design quality factor randomly chosen between 40 and 80. Other than that, the same tests are performed. The results are shown in Table 2. From this table, we find that if we do not know the design

| Final QF | 100 | 90 | 80 | 70 | 60 | 50 |
|---|---|---|---|---|---|---|
| DOQQ | 0 | 0 | 0 | 0 | 0 | .016 |
| CPCD | .088 | .044 | .144 | .132 | .248 | .220 |
| CADC | .004 | 0 | .044 | .104 | .212 | .292 |

**Table 2**. If the design quality factor is unknown, the detection error is higher than previous results, but still sufficiently low. Also, the final JPEG compression plays an important role. As compression becomes lower, the detection becomes less accurate.

quality factor, the detection accuracy becomes lower, as expected. We also find that now the JPEG compression becomes an important factor. As compression becomes more severe, the detection error goes up. This is expected because the compression of images disrupts the artifacts introduced by data hiding, therefore making the hidden content less detectable.

**Mixed data set:** In the previous tests, we build a classifier and then perform detection on images from the same collection. Images from the same collection may have similar content, texture, or processing artifacts, and in real world detection tasks, we do not know which collection the images are from. Therefore, we design this test to partially remove this restriction. We create training and testing sets with images from all three image collections with equal proportion. The results are shown in Table 3

| Final QF | 100 | 90 | 80 | 70 | 60 | 50 |
|---|---|---|---|---|---|---|
| Mixed Set | .001 | .004 | .000 | .001 | .117 | .083 |

**Table 3**. If the training and testing set are created with images from a mixture of three collections, the supervised learning method can still make very accurate detection.

In this test, we found that although we train our classifier and attempt to detect images from mixture of three collections, we still get very accurate prediction at all compression rates. This suggests that the difference between different data collections as well as changes due to hidden data can be learned from a one-step supervised learning.

## 5. CONCLUSION

Our detection-theoretic results for i.i.d. hosts show that the ease with which QIM can be detected depends strongly on the host statistics. Specifically, host PMFs with a sharp peak at the mean change considerably after QIM based hiding, which then becomes easy to detect. This characteristic does hold for typical transform domain image data, which has strong peaks at zero. While the knowledge of host distribution assumed in our detection-theoretic analysis does not hold for image data (where the statistics can vary significantly from image to image), standard supervised learning techniques are shown to perform well. The methods employed here only employ the first-order statistics, and their performance could potentially be further improved by exploiting host memory.

We caution the reader against drawing the conclusion that QIM is inherently easily detectable. The detectability could be reduced by reducing the design level of robustness against attacks, or by reducing the embedding rate. More fundamentally, our work only considers currently proposed QIM schemes, which appear to have been designed with robustness, rather than covertness, in mind. We leave open the issue of whether it is possible to design QIM schemes that are both robust and covert, and point to some recent theoretical results that indicate the potential for such schemes [12].

# References


[1] M.H.M. Costa, "Writing on dirty paper," *IEEE Trans. Info. Theory*, vol. IT-29, no. 3, pp. 439–441, May 1983.

[2] B. Chen and G.W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Info. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[3] J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Trans. on Signal Processing*, vol. 51, no. 4, pp. 1003–1019, 2003.

[4] K. Solanki, N. Jacobsen, S. Chandrasekaran, U. Madhow, and B. S. Manjunath, "High-volume data hiding in images: Introducing perceptual criteria into quantization based embedding," in *Proceedings of ICASSP*, May 2002.

[5] P. Meerwald, "Quantization watermarking in the JPEG2000 coding pipeline," in *5th International Working Conference on Communication and Multimedia Security*, 2001.

[6] P. Bas, N. Le Bihan, and J.-M. Chassery, "Color image watermarking using quaternion Fourier transform," in *Proceedings of ICASSP*, Hong Kong, China, 2003.

[7] K. Li and X.-P. Zhang, "Reliable adaptive watermarking scheme integrated with JPEG2000," in *Proceedings of ISISPA*, Rome, Italy, 2003.

[8] P. Guillon, T. Furon, and P. Duhamel, "Applied public-key steganography," in *Proceedings of SPIE*, San Jose, CA, 2002.

[9] K. Sullivan, O. Dabeer, U. Madhow, B.S. Manjunath, and S. Chandrasekaran, "LLRT based detection of LSB hiding," in *Proceedings of ICIP*, 2003.

[10] S. Lyu and H. Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *5th International Workshop on Information Hiding*, 1999.

[11] V. Poor, *An introduction to signal detection and estimation*, Springer, NY, 1994.

[12] Y. Wang and P. Moulin, "Steganalysis of block-structured stegotext," in *Proceedings of SPIE*, San Jose, CA, 2004.