UNIVERSITY OF CALIFORNIA
Santa Barbara

Image Steganalysis: Hunting & Escaping

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Kenneth Mark Sullivan

Committee in Charge:

Professor Shivkumar Chandrasekaran, Co-Chair

Professor Upamanyu Madhow, Co-Chair

Professor B.S. Manjunath, Co-Chair

Professor Edward J. Delp

Doctor Ramarathnam Venkatesan

September 2005

The Dissertation of
Kenneth Mark Sullivan is approved:

Professor Edward J. Delp

Doctor Ramarathnam Venkatesan

Professor Shivkumar Chandrasekaran, Committee Co-Chairman

Professor Upamanyu Madhow, Committee Co-Chairman

Professor B.S. Manjunath, Committee Co-Chairman

August 2005

Image Steganalysis: Hunting & Escaping

by

Kenneth Mark Sullivan

To the memory of my sister, Kathleen

Acknowledgements

I would like to thank the data hiding troika: Professors Manjunath, Madhow, and Chandrasekaran. Prof. Manjunath taught me how to approach problems and to keep an eye on the big picture. Prof. Madhow has a knack for explaining difficult concepts concisely, and has helped me present my ideas more clearly. Prof. Chandrasekaran always has an interesting new approach to offer, often helping to push my thinking out of local minima. I also would like to think Prof. Delp and Dr. Venkatesan for their time and helpful comments throughout this research.

The research presented here was supported by the Office of Naval Research (ONR #N00014-01-1-0380 and #N00014-05-1-0816), and the Center for Bioimage Informatics at UCSB.

My data hiding colleague, Kaushal Solanki, has been great to work and travel with over the past few years. During my research in the lab I have been lucky to have a bright person in my field to bounce ideas off of and provide sanity checks, literally just a few feet away. Onkar Dabeer was an amazing help, there seems to be little he can not solve.

I will remember more of my years here than just sitting in the lab because of my friends here. John, Tate, Christian, Noah, it's been fun. GTA 100%, Ditch Witchin'...lots of very exciting times occurred.

Jiyun, thanks for serving as my guide in Korea. Ohashi, thanks for your hospitality in Japan. Dmiriti, thanks for translating Russian for me. To the rest of the VRL, past and present: Sitaram, Marco, Baris, Shawn, Jelena, Motaz, Xinding, Thomas, Feddo, and Maurits, I've learned at least as much from lunchtime discussions as I did the rest of the day, I'm going to miss VRL. Judging from the new kids: Nhat, Mary, Mike, and Laura, the future is in good hands.

Additionally, I would like to thank Prof. Ken Rose for providing a space for me in signal compression lab to work in, and to the SCL members over the years: Ashish, Ertem, Jaewoo, Jayanth, Hua, Sang-Uk, Pakpoom (thanks for the ride home!), for making me feel at home there.

I owe a lot to fellow grad students outside my VRL/SCL world. Chowdary, Chin, KGB, Vishi, Rich, Gwen, Suk-seung, thanks for the help and good times.

My friends from back in the day, Dave and Pete, you helped me take much needed breaks from the whole grad school thing.

Finally I would like to thank my family. For the Brust clan, thanks for commiserating with us when Kaeding shanked that field goal. To my aunts Pat and Susan, I am glad to have gotten to know you much better these past few years. My brother Kevin and my parents Mike and Romaine Sullivan have been a constant source of support; I always return from San Diego refreshed.

Curriculum Vitæ

Kenneth Mark Sullivan

Education

2005	Doctor of Philosophy Department of Electrical and Computer Engineering University of California, Santa Barbara.
2002	Master of Science Department of Electrical and Computer Engineering University of California, Santa Barbara.
1998	Bachelor of Science Department of Electrical and Computer Engineering University of California, San Diego

Experience

2001 – 2005	Graduate Research Assistant, University of California, Santa Barbara.
2001, 2005	Teaching Assistant, University of California, Santa Barbara.
1998 – 2000	Hardware/Software Engineer, Tiernan Communications Inc., San Diego.
1997	Intern, TRW Avionics Systems Division, San Diego.

Selected Publications

K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran "Steganalysis for Markov Cover Data with Applications to Images", Submitted to *IEEE Transactions on Information Forensics and Security.*

K. Solanki, K. Sullivan, B. S. Manjunath, U. Madhow, and S. Chandrasekaran, "Statistical Restoration for Robust and Secure Steganography", To appear *Proc. IEEE International Conference on Image Processing (ICIP)*, Genoa, Italy, Sep., 2005.

K. Sullivan, U. Madhow, S. Chandrasekaran and B. S. Manjunath, "Steganalysis of Spread Spectrum Data Hiding Exploiting Cover Memory" In *Proc. IS&T/SPIE's 17th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan. 2005.

O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran and B.S. Manjunath, "Detection of Hiding in the Least Significant Bit", In *IEEE Transactions on Signal Processing, Supplement on Secure Media I*, vol. 52, no. 10, pp. 3046–3058, Oct. 2004.

K. Sullivan, Z. Bi, U. Madhow, S. Chandrasekaran and B.S. Manjunath, "Steganalysis of quantization index modulation data hiding", In *Proc. IEEE International Conference on Image Processing (ICIP)*, Singapore, pp. 1165–1168, Oct. 2004.

K. Sullivan, O. Dabeer, U. Madow, B. S. Manujunath and S. Chandrasekaran "LLRT Based Detection of LSB Hiding" In *Proc. IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, pp. 497–500, Sep. 2003

O. Dabeer, K. Sullivan, U. Madow, S. Chandrasekaran and B. S. Manjunath "Detection of hiding in the least significant bit" In *Proc. Conference on Information Sciences and Systems (CISS)* Mar., 2003.

Abstract

Image Steganalysis: Hunting & Escaping

Kenneth Mark Sullivan

Image steganography, the covert embedding of data into digital pictures, represents a threat to the safeguarding of sensitive information and the gathering of intelligence. Steganalysis, the detection of this hidden information, is an inherently difficult problem and requires a thorough investigation. Conversely, the hider who demands privacy must carefully examine a means to guarantee stealth. A rigorous framework for analysis is required, both from the point of view of the steganalyst and the steganographer. In this dissertation, we lay down a foundation for a thorough analysis of steganography and steganalysis and use this analysis to create practical solutions to the problems of detecting and evading detection. Detection theory, previously employed in disciplines such as communications and signal processing, provides a natural framework for the study of steganalysis, and is the approach we take. With this theory, we make statements on the theoretical detectability of modern steganography schemes, develop tools for steganalysis in a practical scenario, and design and analyze a means of escaping optimal detection.

Under the commonly used assumption of an independent and identically distributed cover, we develop our detection-theoretic framework and apply it to the

steganalysis of LSB and quantization based hiding schemes. Theoretical bounds on detection not available before are derived. To further increase the accuracy of the model, we broaden the framework to include a measure of dependency and apply this expanded framework to spread spectrum and perturbed quantization hiding methods. Experiments over a diverse database of images show our steganalysis to be effective and competitive with the state-of-the-art.

Finally we shift focus to evasion of optimal steganalysis and analyze a method believed to significantly reduce detectability while maintaining robustness. The expected loss of rate incurred is analytically derived and it is shown that a high volume of data can still be hidden.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image steganography, the covert embedding of data into digital pictures, represents a threat to the safeguarding of sensitive information and the gathering of intelligence. Steganalysis, the detection of this hidden information, is an inherently difficult problem and requires a thorough investigation. Conversely, the hider who demands privacy must carefully examine a means to guarantee stealth. A rigorous framework for analysis is required, both from the point of view of the steganalyst and the steganographer.

The main contribution of this work is the development of a foundation for the thorough analysis of steganography and steganalysis and the use of this analysis to create practical solutions to the problems of detecting and evading detection. Image data hiding is a field that lies in the intersection of communications and image processing, so our approach employs elements of both areas. Detection theory, employed in disciplines such as communications and signal processing,

provides a natural framework for the study of steganalysis. Image processing provides the theory and tools necessary to understand the unique characteristics of cover images. Additionally, results from fields such as information theory and pattern recognition are employed to advance the study.

1.1 Data Hiding Background

As long as people have been able to communicate with one another, there has been a desire to do so secretly. Two general approaches to covert exchanges of information have been: communicate in a way understandable by the intended parties, but unintelligible to eavesdroppers; or communicate innocuously, so no extra party bothers to eavesdrop. Naturally both of these methods can be used concurrently to enhance privacy. The formal studies of these methods, cryptography and steganography, have evolved and become increasingly more sophisticated over the centuries to the modern digital age. Methods for hiding data into *cover* or *host* media, such as audio, images, and video, were developed about a decade ago (e.g. [89], [101]). Although the original motivation for the early development of data hiding was to provide a means of "watermarking" media for copyright protection [58], data hiding methods were quickly adapted to steganography [2, 55]. See Figure 1.1 for a schematic of an image steganography system. Although wa-

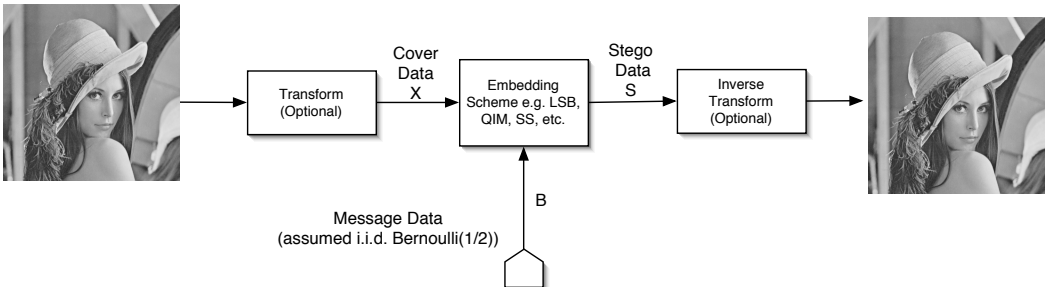

Figure 1.1: Hiding data within an image.

termarking and steganography both imperceptibly hide data into images, they have slightly different goals, and so approaches differ. Watermarking has modest rate requirements, only enough data to identify the owner is required, but the watermark must be able to withstand strong attacks designed to strip it out (e.g. [90], [73]). Steganography generally is subjected to less vicious attacks, however as much data as possible is to be inserted. Additionally, whereas in some cases it may actually serve a watermarker to advertise the existence of hidden data, it is of paramount importance for a steganographer's data to remain hidden. Naturally however, there are those who wish to detect this data. On the heels of developments in steganography come advances in steganalysis, the detection of images carrying hidden data, see Figure 1.2.

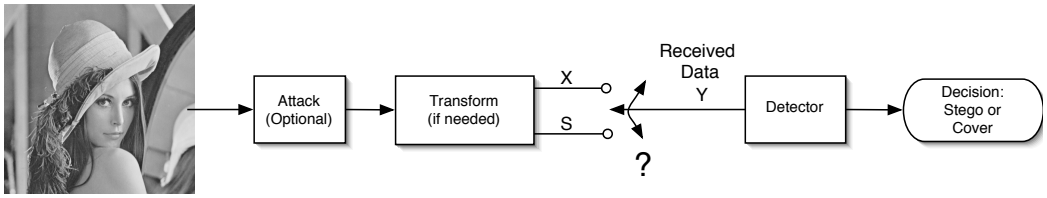

Figure 1.2: Steganalysis flow chart.

1.2 Motivation

The general motivation for steganalysis is to remove the veil of secrecy desired by the hider. Typical uses for steganography are for espionage, industrial or military. A steganalyst may be a company scanning outgoing emails to prevent the leaking of proprietary information, or an intelligence gatherer hoping to detect communication between adversaries.

Steganalysis is an inherently difficult problem. The original cover is not available, the number of steganography tools is large, and each tool may have many tunable parameters. However because of the importance of the problem there have been many approaches. Typically an intuition on the characteristics of cover images is used to determine a decision statistic that captures the effect of data hiding and allow discrimination between natural images and those containing hidden data. The question of the optimality of the statistic used is generally left unanswered. Additionally, the question of how to calibrate these statistics is also left open. We have therefore seen an iterative process of steganography and

steganalysis: a steganographic method is detected by a steganalysis tool, a new steganographic method is invented to prevent detection, which in turn is found to be susceptible to an improved steganalysis. It is not known then what the limits of steganalysis are, an important question for both the steganographer and steganalyst. It is hoped by careful analysis that some measure of optimal detection can be obtained.

1.3 Main Contributions

- **Detection-theoretic Framework.** Detection theory is well-developed and is naturally suited to the steganalysis problem. We develop a detection-theoretic approach to steganalysis general enough to estimate the performance of theoretically optimal detection yet detailed enough to help guide the creation of practical detection tools [21, 85, 20].

- **Practical Detection of Hiding Methods.** In practice, not enough information is available to use optimal detection methods. By devising methods of estimating this information from either the received data, or through supervised learning, we created methods that practically detect three general classes of data hiding: least significant bit (LSB) [21, 85, 20], quantization

index modulation (QIM) [84], and spread spectrum (SS) [87, 86]. These methods compare favorably with published detection schemes.

- **Expand Detection-theoretic Approach to Include Dependencies.** Typically analysis of the steganalysis problem has used an independent and identically distributed (i.i.d.) assumption. For practical hiding media, this assumption is too simple. We take the next logical step and augment the analysis by including Markov chain data, adding statistically dependent data to the detection-theoretic approach [87, 86].

- **Evasion of Optimal Steganalysis.** From our work on optimal steganalysis, we have learned what is required to escape detection. We use our framework to guide evasion efforts and successfully reduce the effectiveness of previously successful detection for dithered QIM [82]. This analysis is also used to derive a formulation of the rate of secure hiding for arbitrary cover distributions.

1.4 Notation, Focus, and Organization

We refer to original media with no hidden data as *cover* media, and media containing hidden data as *stego* media (e.g. cover images, stego transform coefficients). The terms *hiding* or *embedding* are used to denote the process of

adding hidden data to an image. We use the term *robust* to denote the ability of a data hiding scheme to withstand changes incurred to the image between the sender and intended receiver. These changes may be from a malicious attack, transmission noise, or common image processing transformations, most notably compression. By *detection*, we mean that a steganalyst has correctly classified a stego image as containing hidden data. *Decoding* is used to denote the reception of information by the intended receiver. We use *secure* in the steganographic sense, meaning safe from detection by steganalysis. We use capital letters to denote a random variable, and lower case letters to denote the value of its realization. Boldface indicates vectors (lower case) and matrices (upper case). For probability mass functions we use either vector/matrix notation: $\mathbf{p}^{(X)} : p_i^{(X)} = P(X = i)$, $\mathbf{M}_{ij}^{(X)} = P(X_1 = i, X_2 = j)$ or function notation: $P_X(x) = P(X = x)$, $P_{X_1,X_2}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$ where context determines which is more convenient. A complete list of symbols and acronyms used is provided in the Appendix.

Classification between cover and stego is often referred to as "passive" steganalysis while extracting hidden information is referred to as "active" steganalysis. Extraction can also be used as an attack on a watermarking system: if the watermark is known, it can easily be removed without distorting the cover image. In most cases, the extraction is actually a special case of cryptanalysis (e.g. [62]),

a mature field in its own right. We focus exclusively on passive steganalysis and drop the term "passive" where clear. To confuse matters, the literature also often refers to a "passive" and "active" warden. In both cases, the warden controls the channel between the sender and receiver. A passive warden lets an image pass through unchanged if it is judged to not contain hidden data. An active warden attempts to destroy any possible hidden data by making small changes to the image, similar in spirit to a copyright violator attempting to remove a watermark. We generally focus on the passive warden scenario, since many aspects of the active warden case are well studied in watermarking research. However, we discuss the robustness of various hiding methods to an active warden and other possible attacks/noise.

Furthermore, though data hiding techniques have been developed for audio, image, video, and even non-multimedia data sources such as software [91], we focus on digital images. Digital images are well suited to data hiding for a number of reasons. Images are ubiquitous on the Internet; posting an image on a website or attaching a picture to an email attracts no attention. Even with modern compression techniques, images are still relatively large and can be changed imperceptibly, both important for covert communication. Finally there exist several well-developed methods for image steganography, more than for any other data hiding medium. We focus on grayscale images in particular.

To provide context for our examination of steganalysis, in the following chapter we review steganography and steganalysis research presented in the literature. In Chapter 3, we explain the detection-theoretic framework we use throughout the study, and apply it to the steganalysis of LSB and QIM hiding schemes. In Chapter 4, we broaden the framework to include a measure of dependency and apply this expanded framework to SS and PQ hiding methods. In Chapter 5, we shift focus to evasion of optimal steganalysis and analyze a method believed to significantly reduce detectability while maintaining adequate rate and robustness. We summarize our conclusions and discuss future research directions in Chapter 6.

Chapter 2

Steganography and Steganalysis

We here survey the concurrent development of image steganography and steganalysis. Research and development of steganography preceded steganalysis, and steganalysis has been forced to catch up. More recently, steganalysis has had some success and steganographers have had to more carefully consider the stealthiness of their hiding methods.

2.1 Basic Steganography

Digital image steganography grew out of advances in digital watermarking. Two early watermarking methods which became two early steganographic methods are: overwriting the least significant bit (LSB) plane of an image with a message; and adding a message bearing signal to the image [89].

The LSB hiding method has the advantage of simplicity of encoding, and a guaranteed successful decoding if the image is unchanged by noise or attack. How-

ever the LSB method is very fragile to any attack, noise, or even standard image processing such as compression [52]. Additionally, because the least significant bit plane is overwritten, the data is irrecoverably lost. For the steganographer, however, there are many scenarios with which the image remains untouched, and the cover image can be considered disposable. As such, LSB hiding is still very popular today; a perusal of tools readily available online reveals numerous LSB embedding software packages [74]. We examine LSB hiding in greater detail in Chapter 3.

The basic idea of additive hiding is straightforward. Typically the binary message modulates a sequence known by both encoder and decoder, and this is added to the image. This simplicity lends itself to adaptive improvements. In particular, unlike LSB, additive hiding schemes can be designed to withstand changes to the image such as JPEG compression and noise [101]. Additionally, if the decoder correctly receives the message, he or she can simply subtract out the message sequence, recovering the original image (assuming no noise or attack). Much watermarking research then has focused on additive hiding schemes, specifically improving robustness to malicious attacks (e.g. [73],[90]) deliberately designed to remove the watermark.

A commonly used adaptation of the additive hiding scheme is the spread spectrum (SS) method introduced by Cox et al [19]. As suggested by the name,

the message is spread (whitened) as is typically done in many applications such as wireless communications and anti-jam systems [66], and then added to the cover. This method, with various adaptations, can be made robust to typical geometric and noise adding attacks. Naturally newer attacks are created (e.g. [62]) and new solutions to the attacks are proposed. As with LSB hiding, spread spectrum and close variants are also used for steganography [60, 31]. We describe SS hiding in greater detail in Chapter 4.

An inherent problem with SS hiding, and any additive hiding, is interference from the cover medium. This interference can cause errors at the decoder, or equivalently, lowers the amount of data that can be accurately received. However, the hider has perfect knowledge of the interfering cover; surely the channel has a higher capacity than if the interference were unknown. Work done by Gel'Fand and Pinsker [39], as well as Costa [17], on hiding in a channel with side information known only by the encoder show that the capacity is not effected by the known noise at all. In other words, if the data is encoded correctly by the hider, there is effectively no interference from the cover, and the decoder only needs to worry about outside noise or attacks. The encoder used by Costa for his proof is not readily applicable. However, for the data hiding problem, Chen and Wornell proposed quantization index modulation QIM [14] to avoid cover interference. This coding method and its variants achieve, or closely achieve, the capacity

predicted by Costa. The basic idea is to hide the message data into the cover by quantizing the cover with a choice of quantizer determined by the message. The simplest example is so-called odd/even embedding. With this scheme, a continuous valued cover sample is used to embed a single bit. To embed a 0, the cover sample is rounded to the nearest even integer, to embed a 1, round to the nearest odd number. The decoder, with no knowledge of the cover, can decode the message so long as perturbations (from noise or attack) do not change the values by more than 0.5. Other similar approaches have been proposed such as the scalar Costa scheme (SCS) by Eggers et al [25]. This class of embedding techniques is sometimes referred to as quantization-based techniques, dirty paper codes (from the title of Costa's paper), and binning methods [104]; we use the term QIM. As the expected capacity is higher than the host interference case, QIM is well suited for steganographic methods [81, 54]. This hiding technique in described in greater detail in Chapter 3.

All of the above methods can be performed in the spatial domain (i.e. pixel values) or in some transform domain. Popular transforms include the two-dimensional discrete cosine transform (DCT), discrete Fourier transform (DFT) [50] and discrete wavelet transforms (DWT) [92]. These transforms may be performed blockwise, or over the entire image. For a blockwise transform, the image is broken into smaller blocks (8×8 and 16×16 are two popular sizes), and the transform

is performed individually on each block. The advantage of using transforms is that it is generally easier to balance distortion introduced by hiding and robustness to noise or attack in the transform domain then in the pixel domain. These transforms can in principle be used with any hiding scheme. LSB hiding however requires digitized data, so continuous valued transform coefficients must be quantized. Transform LSB hiding is therefore generally limited to compressed (with JPEG [94] for example) images, in which the transform coefficients are quantized. Additionally, QIM has historically been used much more often in the transform domain.

We have then three main categories of hiding methods: LSB, SS, and QIM. Data hiding is an active field with new methods constantly introduced, and certainly some of these do not fit into these three categories. However the three we focus on are the most commonly used today, and provide a natural starting point for study. In addition to immediately applicable results, it is hoped that the analysis of these schemes yields findings adaptable to future developments. We now examine some of the steganalysis methods introduced over the last decade to detect these schemes, particularly the popular LSB method. Steganography research has not been idle, and we also review the hider's response to steganalysis.

2.2 Steganalysis

There is a myriad of approaches to the steganalysis problem. Since the general steganalysis problem, discriminating between images with hidden data and images without, is very broad, some assumptions are made to obtain a well-posed problem. Typically these assumptions are made on the cover data, the hiding method, or both. Each steganalysis method presented here uses a different set of assumptions; we look at the advantages and disadvantages of these various approaches.

2.2.1 Detecting LSB Hiding

An early method used to detect LSB hiding is the χ^2 (chi-squared) technique [100], later successfully used by Provos' stegdetect [69] for detection of LSB hiding in JPEG coefficients. We first note that generally the binary message data is assumed to be i.i.d. with the probability of 0 equality to the probability of 1. If the hider's intended message does not have these properties, a wise steganographer would use an entropy coder to reduce the size of the message; the compressed version of the message should fulfill the assumptions. Because 0 and 1 are equally likely, after overwriting the LSB, it is expected that the number of pixels in a pair of values which share all but the LSB are equalized, see Figure 2.1. Although

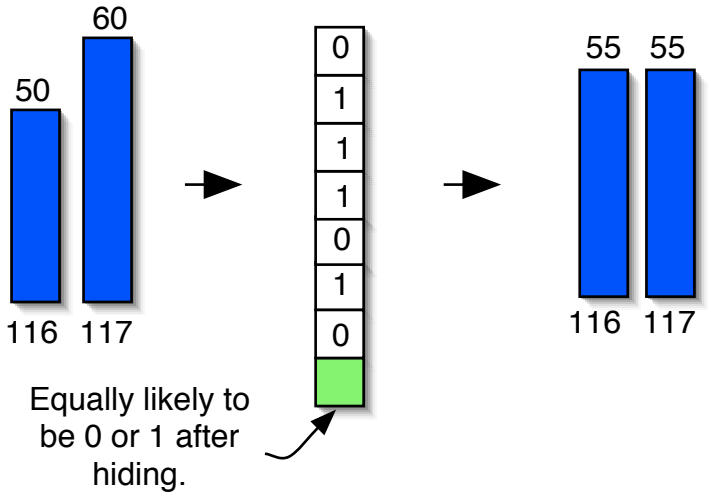

Figure 2.1: Hiding in the least significant bit tends to equalize adjacent histogram bins that share all other bits. In this example of hiding in 8-bit values, the number of pixels with grayscale value 116 becomes equal to the number with value 117.

we would expect these numbers to be close before hiding, we do not expect them to be equal in typical cover data. Due to this effect, if a histogram of the stego data is taken over all pixel values (e.g. 0 to 255 for 8-bit data), a clear "step-like" trend can be seen. We know then exactly what the histogram is expected to look like after LSB hiding in every pixel (or DCT coefficient). The χ^2 test is a goodness-of-fit measure which analyzes how close the histogram of the image under scrutiny is to the expected histogram of that image with embedded data. If it is "close", we decide it has hidden data, otherwise not. In other words, χ^2 is a measure of the likelihood that the unknown image is stego. An advantage of this is that no knowledge of the original cover histogram is required. However a

weakness of the χ^2 test is that it only says how likely the received data is stego, it does not say how likely it is cover. A better test is to decide if it is closer to stego than to cover, otherwise an arbitrary choice must be made as to when it is far enough to be considered clean. We explore the cost of this more fully in Chapter 3. In practice the χ^2 test works reasonably well in discriminating between cover and stego. The χ^2 is an example of an early approach to detecting changes using the statistics of an image, in this case using an estimate of the probability distribution, i.e. a histogram. Previous detection methods were often visual, i.e. for some hiding methods it was found that, in some domain, the hiding was actually recognizable by the naked eye. Visual attacks are easily compensated for, but statistical detection is more difficult to thwart.

Another LSB detection scheme was proposed by Avcibas et al [4] using binary similarity measures between the 7th bit plane and the 8th (least significant) bit plane. It is assumed that there is a natural correlation between the bit planes that is disrupted by LSB hiding. This scheme does not auto-calibrate on a per image basis, and instead calibrates on a training set of cover and stego images. The scheme works better than a generic steganalysis scheme, but not as well as state-of-the-art LSB steganalysis.

Two more recent and powerful LSB detection methods are the RS (regular/singular) scheme [33] and the related sample pair analysis [24]. The RS

scheme, proposed by Fridrich et al, is a specific steganalysis method for detecting LSB data hiding in images. Sample pair analysis is a more rigorous analysis due to Dumitrescu et al of the basis of the RS method, explaining why and when it works. The sample pairs are any pair of values (not necessarily consecutive) in a received sequence. These pairs are partitioned into subsets depending on the relation of the two values to one another. Is is assumed that in a cover image the number of pairs in each subset are roughly equal. It is shown that LSB hiding performs a different function on each subset, and so the number of pairs in the subsets are not equal. The amount of disruption can be measured and related to the known effect of LSB hiding to estimate the rate of hiding. Although the initial assumption does not require interpixel dependencies, it can be shown that correlated data provides stronger estimates than uncorrelated data. The RS scheme, a practical detector of LSB data hiding, uses the same basic principle as sample pair analysis. As in sample pair analysis, the RS scheme counts the number of occurrences of pairs in given sets. The relevant sets, regular and singular (hence RS), are related to but slightly different from the sets used in sample pair analysis. Also as in sample pair analysis, equations are derived to estimate the length of hidden messages. Since RS employs the same principle as sample pair analysis, we would expect it to also work better for correlated cover data. Indeed the RS scheme focuses on spatially adjacent image pixels, which are known to be highly

correlated. In practice RS analysis and sample pair analysis perform comparably. Recently Roue et al [72] use estimates of the joint probability mass function (PMF) to increase the detection rate of RS/sample pair analysis. We explore the joint PMF estimate in greater detail in Chapter 4. A recent scheme, also by Fridrich and Goljan [32], uses local estimators based on pixel neighborhoods to slightly improve LSB detection over RS.

2.2.2 Detecting Other Hiding Methods

Though most of the focus of steganalysis has been on detecting LSB hiding, other methods have also been investigated.

Harmsen and Pearlman studied [45] the steganalysis of additive hiding schemes such as spread spectrum. Their decision statistic is based initially on a PMF estimate, i.e. a histogram. Since additive hiding is an addition of two random variables: the cover and the message sequence, the PMF of cover and message sequences are convolved. In the Fourier domain, this is equivalent to multiplication. Therefore the DFT of the histogram, termed the histogram characteristic function (HCF), is taken. It is shown for typical cover distributions that the expected value, or center of mass (COM), of the HCF does not increase after hiding, and in practice typically decreases. The authors choose then to use the COM as a feature to train a Bayesian multivariate classifier to discriminate between cover

and stego. They perform tests on RGB images, using a combined COM of each color plane, with reasonable success in detecting additive hiding.

Celik et al [11] proposed using rate-distortion curves for detection of LSB hiding and Fridrich's content-independent stochastic modulation [31] which, as studied here, is statistically identical to spread spectrum. They observe that data embedding typically increases the image entropy, while attempting to avoid introducing perceptual distortion to the image. On the other hand, compression is designed to reduce the entropy of an image while also not inducing any perceptual changes. It is expected therefore that the difference between a stego image and its compressed version is greater than the difference between a cover and its compressed form. Distortion metrics such as mean squared error, mean absolute error, and weighted MSE are used to measure the difference between an image and compressed version of the image. A feature vector consisting of these distortion metrics for several different compression rates (using JPEG2000) is used to train a classifier. False alarm and missed detection rates are each about 18%.

2.2.3 Generic Steganalysis: Notion of Naturalness

The following schemes are designed to detect any arbitrary scheme. For example, rather than classifying between cover images and images with LSB hiding, they discriminate between cover images and stego images with any hiding scheme,

or class of hiding schemes. The underlying assumption is that cover images posses some measurable naturalness that is disrupted by adding data. In some respects this assumption lies at the heart of all steganalysis. To calibrate the features chosen to measure "naturalness", the systems learn using some form of supervised training.

An early approach was proposed by Avcibas et al [3, 5], to detect arbitrary hiding schemes. Avcibas et al design a feature set based on image quality metrics (IQM), metrics designed to mimic the human visual system (HVS). In particular they measure the difference between a received image and a filtered (weighted sum of 3×3 neighborhood) version of the image. This is very similar in spirit to the work by Celik et al, except with filtering instead of compression. The key observation is that filtering an image without hidden data changes the IQMs differently than an image with hidden data. The reasoning here is that the embedding is done locally (either pixel-wise or blockwise), causing localized discrepancies. We see these discrepancies exploited in many steganalysis schemes. Although their framework is for arbitrary hiding, they also attempted to fine tune the choice of IQMs for two classes of embedding schemes: those designed to withstand malicious attack, and those not. A multivariate regression classifier is trained with examples of images with and without hidden data. This work is an early example of supervised learning in steganalysis. Supervised learning is used to overcome

the steganalyst's lack of knowledge of cover statistics. From experiments performed, we note that there is a cost for generality: the detection performance is not as powerful as schemes designed for one hiding scheme. The results however are better than random guessing, reinforcing the hypothesis of the inherent "unnaturalness" of data hiding.

Another example of using supervised learning to detect general steganalysis is the work of Lyu and Farid [57, 56, 28]. Lyu and Farid use a feature set based on higher-order statistics of wavelet subband coefficients for generic detection. The earlier work used a two-class classifier to discriminate between cover and stego images made with one specific hiding scheme. Later work however uses a one-class, multiple hypersphere, support vector machine (SVM) classifier. The single class is trained to cluster clean cover images. Any image with a feature set falling outside of this class is classified as stego. In this way, the same classifier can be used for many different embedding schemes. The one-class cluster of feature vectors can be said to capture a "natural" image feature set. As with Avcibas et al's work, the general applicability leads to a performance hit in detection power compared with detectors tuned to a specific embedding scheme. However the results are acceptable for many applications. For example, in detecting a range of different embedding schemes, the classifier has a miss probability between 30-40% for a false alarm rate around 1% [57]. By choosing the number of hyperspheres

used in the classifier, a rough tradeoff can be made between false alarms and misses.

Martin et al [59] attempt to directly use the notion of the "naturalness" of images to detect hidden data. Though they found that data hidden certainly caused shifts from the natural set, knowledge of the specific data hiding scheme provides far better detection performance.

Fridrich [30] presented another supervised learning method tuned to JPEG hiding schemes. The feature vector is based on a variety of statistics of both spatial and DCT values. The performance seems to improve over previous generic detection schemes by focusing on a class of hiding schemes [53].

From all of these approaches, we see that generalized detection is possible, confirming that data hiding indeed fundamentally perturbs images. However, as one would expect, in all cases performance is improved by reducing the scope of detection. A detector tuned to one hiding scheme performs better than a detector designed for a class of schemes, which in turn beats general steganalysis of all schemes.

2.2.4 Evading Steganalysis

Due to the success of steganalysis in detecting early schemes, new steganographic methods have been invented in an attempt to evade detection.

F5 by Westfeld [99] is a hiding scheme that changes the LSB of JPEG coefficients, but not by simple overwriting. By increasing and decreasing coefficients by one, the frequency equalization noted in standard LSB hiding is avoided. That is, instead of standard LSB hiding, where an even number is either unchanged or increased by one, and an odd is either unchanged or *decreased* by one, both odd and even numbers are increased and decreased. This method does indeed prevent detection by the χ^2 test. However Fridrich et al [35] note that although F5 hiding eliminates the characteristic "step-like" histogram of standard LSB hiding, it still changes the histogram enough to be detectable. A key element in their detection of F5 is the ability to estimate the cover histogram. As mentioned above, the χ^2 test only estimates the likelihood of an image being stego, providing no idea of how close it is to cover. By estimating the cover histogram, an unknown image can be compared to both an estimate of the cover, and the expected stego, and whichever is closest is chosen. Additionally, by comparing the relative position of the unknown histogram to estimates of cover and stego, an estimate of the amount of data hidden, the hiding rate, can be determined. The method of estimating the cover histogram is to decompress, crop the image by 4 pixels (half a JPEG block), and recompress with the same quantization matrix (quality level) as before. They find this cropped and recompressed image is statistically very close to the original, and generalize this method to detection of other JPEG hiding schemes [36]. We

note that detection results are good, but a quadratic distance function between the histograms is used, which is not in general the optimal measure [67, 105]. Results may be further improved by a more systematic application of detection theory.

Another steganographic scheme based on LSB hiding, but designed to evade the χ^2 test is Provos' Outguess 0.2b [68]. Here LSB hiding is done as usual (again in JPEG coefficients), but only half the available coefficients are used. The remaining coefficients are used to compensate for the hiding, by repairing the histogram to match the cover. Although the rate is lower than F5 hiding, since half the coefficients are not used, we would expect this to not only be undetectable by χ^2, but by Fridrich's F5 detector, and in fact by *any* detector using histogram statistics. However, because the embedding is done in the blockwise transform domain, there are changes in the spatial domain at the block borders. Specifically, the change to the spatial joint statistics, i.e. the dependencies between pixels, is different than for standard JPEG compression. Fridrich et al are able to exploit these changes at the JPEG block boundaries [34]. Again using a decompress-crop-recompress method of estimating the cover (joint) statistics, they are able to detect Outguess and estimate the message size with reasonable accuracy. We analyze the use of interpixel dependencies for steganalysis in Chapter 4. In a similar vein, Wang and Moulin [97], analyze detecting block-DCT based spread-

spectrum steganography. It is assumed that the cover is stationary, and so the interpixel correlation should be the same for any pair of pixels. Two random variables are compared: the difference in values for pairs of pixels straddling block borders, and the difference of pairs within the block. Under the cover stationarity assumption these should have the same distribution, i.e. the difference histogram should be the same for border pixels and interior pixels. A goodness-of-fit measure is used to test the likelihood of that assumption on a received image. As with the χ^2 goodness-of-fit test, the threshold for deciding data is hidden varies from image to image.

A method that attempts to not only preserve the JPEG coefficient histogram but also interpixel dependencies after LSB hiding is presented by Franz [29]. To preserve the histogram, the message data distribution is matched to that of the cover data. Recall that LSB hiding tends to equalize adjacent histogram bins because the message data is equally likely to be 0 or 1. If however the imbalance between adjacent histogram bins is mimicked by the message data, the hiding does not change the histogram. Unfortunately this increase in security does not come for free. As mentioned earlier, compressed message data has equal probabilities of 0 and 1. This is the maximum entropy distribution for binary data, meaning the most information is conveyed by the data. Binary data with unequal probabilities of 0 and 1 carries less information. Thus, if a message is converted to

match the cover histogram imbalance, the number of bits hidden must increase. The maximum effective hiding rate is the entropy: $H_b(p) = -p\log_2(p) - (1-p)\log_2(1-p)$, where p is the probability of 0 [18]. To decrease detection of changes to dependencies, the author suggests only embedding in pairs of values that are independent. A co-occurrence matrix, a two-dimensional histogram of pixel pairs, is used to determine independence. Certainly not all values are independent but the author shows the average loss of capacity is only about 40%, which may be an acceptable loss to ensure privacy. It is not clear though how a receiver can be certain which coefficients have data hidden, or if similar privacy can be found for less loss of capacity. This method is detected by Böhme and Westfeld [8] by exploiting the asymmetric embedding process. That is, by not embedding in some values due to their dependencies, a characteristic signature is left in the co-occurrence matrix. We show in Chapter 4 that under certain assumptions the co-occurrence matrix is the basis for optimal statistical detection.

Eggers et al [26] suggest a method of data-mappings that preserve the first-order statistics, called histogram-preserving data-mapping (HPDM). As with the method proposed by Franz, the distribution of the message is designed to match the cover, resulting in a loss of rate. Experiments show this reduces the Kullback-Leibler divergence between the cover and stego distributions, and thus reduces the probabilty of detection (more on this below). Since only the histogram is

matched, Lyu and Farid's higher-order statistics learning algorithm is able to detect it. Tzschoppe et al [88] suggest a minor modification to avoid detection: basically not hiding in perceptually significant values. We investigate a means to match the histogram exactly, rather than on average, while also preserving perceptually significant values, in Chapter 5.

Fridrich and Goljan [31] propose the stochastic modulation hiding scheme designed to mimic noise expected in an image. The non-content dependent version allows arbitrarily distributed noise to be used for carrying the message. If Gaussian noise is used, the hiding is statistically the same as spread spectrum, though with a higher rate than typical implementations. The content dependent version adapts the strength of the hiding to the image region. As statistical tests typically assume one statistical model throughout the image, content adaptive hiding may evade these tests by exploiting the non-stationarity of real images.

General methods for adapting hiding to the cover face problems with decoding. The intended receiver may face ambiguities over where data is and is not hidden. Coding frameworks for overcoming this problem have been presented by Solanki et al [81] for a decoder with incomplete information on hiding locations and by Fridrich et al [38] when the decoder has no information. This allows greater flexibility in designing steganography to evade detection.

To escape RS steganalysis, Yu et al propose an LSB scheme designed to resist detection from both χ^2 and RS tests [103]. As in F5, the LSB is increased or decreased by one with no regard to the value of the cover sample. Additionally some values are reserved to correct the RS statistic at the end. Since the embedding is done in the spatial domain, rather than in JPEG coefficients, Fridrich et al's F5 detector [35] is not applicable, though it is not verified that other histogram detection methods would not work. Experiments are performed showing the method can foil RS and χ^2 steganalysis.

2.2.5 Detection-Theoretic Analysis

We have seen many cases of a new steganographic scheme created to evade current steganalysis. In turn this new scheme is detected by an improved detector, and steganographers attempts to thwart the improved detector. Ideally, instead of iterating in this manner, the inherent detectability of a steganographic scheme to *any* detector, now or in the future, could be pre-determined. An approach that yields hope of determining this is to model an image as a realization of a random process, and leverage detection theory to determine optimal solutions and estimate performance. The key advantage of this model for steganalysis is the availability of results prescribing *optimal* (error minimizing) detection methods as well as providing estimates of the results of optimal detection. Additionally the

study of idealized detection often suggests an approach for practical realizations. There has been some work with this approach, particularly in the last couple of years.

An early example of a detection-theoretic approach to steganalysis is Cachin's work [10]. The steganalysis problem is framed as a hypothesis test between cover and stego hypotheses. Cachin suggests a bound on the Kullback-Leibler (K-L) divergence (relative entropy) between the cover and stego distributions as a measure of the security between cover and stego. This security measure is denoted ϵ-secure, where ϵ is the bound on the K-L divergence. If ϵ is zero, the system is described as perfectly secure. Under an i.i.d. assumption, by Stein's Lemma [18] this is equivalent to bounds on the error rates of an optimal detector. We explore this reasoning in greater detail in Chapter 3.

Another information theoretic derivation is done for a slightly different model by Zölner et al [107]. They first assume that the steganalyst has access to the exact cover, and prove the intuition that this can never be made secure. They modify the model so that the detector has some, but not complete, information on the cover. From this model they find constraints on conditional entropy similar to Cachin's, though more abstract and hence more difficult to evaluate in practice.

Chandramouli and Memon [13] use a detection-theoretic framework to analyze LSB detection. However, though the analysis is correct, the model is not accurate

enough to provide practical results. The cover is assumed to be a zero mean white Gaussian, a common approach. Since LSB hiding effectively either adds one, subtracts one, or does nothing, they frame LSB hiding as additive noise. If it seems likely that the data came from a zero mean Gaussian, it is declared cover. If it seems likely to have come from a Gaussian with mean of one or minus one, it is declared stego. However, the hypothesis source distribution depends on the current value. For example, the probability that a four is generated by LSB hiding is the probability the message data was zero and the cover was either four or five; so the stego likelihood is half the probability of either a four or five occurring from a zero mean Gaussian. Under their model however, if a four is received, the stego hypothesis distributions are a one mean Gaussian and a negative one mean Gaussian. We present a more accurate model of LSB detection in Chapter 3.

Guillon et al [43] analyze the detectability of QIM steganography, and observe that QIM hiding in a uniformly distributed cover does not change the statistics. That is, the stego distribution is also uniform, and the system has $\epsilon = 0$. Since typical cover data is not in fact uniformly distributed, they suggest using a non-linear "compressor" to convert the cover data to a uniformly distributed intermediate cover. The data is hidden into this intermediate cover with standard QIM, and then the inverse of the function is used to convert to final stego

data. However Wang and Moulin [98] point out that such processing may be unrealizable.

Using detection theory from the steganographer's view point, Sallee [75] proposed a means of evading optimal detection. The basic idea is to create stego data with the same distribution model as the cover data. That is, rather than attempting to mimic the exact cover distribution, mimic a parameterized model. The justification for this is that the steganalyst does not have access to the original cover distribution, but must instead use a model. As long as the steganographer matches the model the steganalyst is using, the hidden data does not look suspicious. The degree with which the model can be approximated with hidden data can be described as ϵ-secure with respect to that model. A specific method for hiding in JPEG coefficients using a Cauchy distribution model is proposed. Though this specific method is found to be vulnerable by Böhme and Westfeld [7], the authors stress their successful detection is due to a weakness in the model, rather than the general framework. More recently Sallee has included [76] a defense against the blockiness detector [34], by explicitly compensating the blockiness measure after hiding with unused coefficients, similar to OutGuess' histogram compensation. The author concedes an optimal solution would require a method of matching the complete joint distribution in the pixel domain, and leaves the development of this method to future work.

A thorough detection-theoretic analysis of steganography was recently presented by Wang and Moulin [98]. Although the emphasis is on steganalysis of block-based schemes, they make general observations of the detectability of SS and QIM. It is shown for Gaussian covers that spread spectrum hiding can be made to have zero divergence ($\epsilon = 0$). However it is not clear if this extends to arbitrary distributions, and additionally requires the receiver to know the cover distribution, which is not typically assumed for steganography. It is shown that QIM generally is not secure. They suggest alternative hiding schemes that can achieve zero divergence under certain assumptions, though the effect on the rate of hiding and robustness is not immediately transparent. Moulin and Wang address the secure hiding rate in [63], and derive a information theoretic capacity for secure hiding for a specified cover distribution and distortion constraints on hider and attacker. The capacity is explicitly derived for a Bernoulli(1/2) (coin toss) cover distribution and Hamming distance distortion constraint, and capacity achieving codes are derived. However for more complex cover distributions and distortion constraints, the derivation of capacity is not at all trivial. We analyze a QIM scheme empirically designed for zero divergence and derive the expected rate and robustness in Chapter 5.

More recently, Sidorov [78] presented work done on using hidden Markov model (HMM) theory for the study of steganalysis. He presents analysis on using Markov

chain and Markov random field models, specifically for detection of LSB. Though the framework has great potential, the results reported are sparse. He found that a Markov chain (MC) model provided poor results for LSB hiding in all but high-quality or synthetic images, and suggested a Markov random field (MRF) model, citing the effectiveness of the RS/sample pair scheme. We examine Markov models and steganalysis in Chapter 4.

Another recent paper applying detection theory to steganalysis is Hogan et al's QIM steganalysis [46]. Statistically optimal detectors for several variants of QIM are derived, and experimental results found. The results are compared to Farid's general steganalysis detector [28], and not surprisingly are much better. We show their results are consistent with our findings on optimal detection of QIM in Chapter 3.

2.3 Summary

There is a great deal to learn from the research presented over the years. We review the lessons learned and note how they apply to our work.

We have seen in many cases a new steganographic scheme created to evade current steganalysis which in turn is detected by an improved detector. Ideally, instead of iterating in this manner, the inherent detectability of a steganographic

scheme to *any* detector, now or in the future, could be pre-determined. The detection-theoretic framework we use to attempt this is presented in Chapter 3

Not surprisingly, detecting many steganalysis schemes at once is more difficult than detecting one method at a time. We use a general framework, but approach each hiding scheme one at a time. LSB hiding is a natural starting point, and we begin our study of steganalysis there. Other hiding methods have received less attention, hence we continue our study with QIM, SS, and PQ, a version of QIM adapted to reduce detectablity [38].

Under an i.i.d. model, the marginal statistics, i.e., frequency of occurrence or histogram, are sufficient for optimal detection. However, we have seen that schemes based on marginal statistics are not as powerful as schemes exploiting interpixel correlations in some way. A natural next step then is to broaden the model to account for interpixel dependencies. We extend our detection-theoretic framework to include a measure of dependency in Chapter 4.

We note that a common solution to the lack of cover statistic information, that is, the problem of how to calibrate the decision statistic, is to use some form of supervised learning [30, 57, 5, 11, 45, 4]. Since this seems to yield reasonable results, we often turn to supervised learning when designing practical detectors.

Chapter 3

Detection-theoretic Approach to Steganalysis

In this chapter we introduce the detection-theoretic approach that we use to analyze steganography, and to develop steganalysis tools. We relate the theory to the steganalysis problem, and establish our general method. This approach is applied to the detection of least significant bit (LSB) hiding and quantization index modulation (QIM), under an assumption of i.i.d. cover data. Both the limits of idealized optimal detection are found as well as tools for detection under realistic scenarios.

3.1 Detection-theoretic Steganalysis

As mentioned in Chapter 2, a systematic approach to the study of steganalysis is to model an image as a realization of a random process, and to leverage detection

theory to determine optimal solutions and to estimate performance. Detection theory is well developed and has been applied to a variety of fields and applications [67]. Its key advantage for steganalysis is the availability of results prescribing *optimal* (error minimizing) detection methods as well as providing estimates of the results of optimal detection.

The essence of this approach is to determine which random process generated an unknown image under scrutiny. It is assumed that the statistics of cover images are different than the statistics of a stego image. The statistics of samples of a random process are completely described by the joint probability distributions: the probability density function (pdf) for a continuous-valued random process and by the probability mass function (PMF) for a discrete-valued random process. With the distribution, we can evaluate the probability of any event.

Steganalysis can be framed as a hypothesis test between two hypotheses: the null hypothesis H_0, that the image under scrutiny is a clean cover image, and H_1, the stego hypothesis, that the image has data hidden in it. The steganalyst uses a detector to classify the data samples of an unknown image into one of the two hypotheses. Let the observed data samples, that is, the elements of the image under scrutiny, be denoted as $\{Y_n\}_{n=1}^{N}$, where Y_n take values in an alphabet $\mathcal{Y}$. Mathematically, a detector δ is characterized by the acceptance region $A \in \mathcal{Y}^N$

of hypothesis H_0:

$$\delta(Y_1, \ldots, Y_N) = \begin{cases} H_0 & \text{if } (Y_1, \ldots, Y_N) \in A, \\ H_1 & \text{if } (Y_1, \ldots, Y_N) \in A^c. \end{cases}$$

In steganalysis, before receiving any data, the probabilities $P(H_0)$ and $P(H_1)$ are unknown; who knows how many steganographers exist? In the absence of this a priori information, we use the Neyman-Pearson formulation of the optimal detection problem: for $\alpha > 0$ given, minimize

$$P\left(\text{Miss}\right) = P\left(\delta(Y_1, \ldots, Y_N) = H_0 | H_1\right)$$

over detectors δ which satisfy

$$P\left(\text{False alarm}\right) = P\left(\delta(Y_1, \ldots, Y_N) = H_1 | H_0\right) \leq \alpha.$$

In other words, minimize the probability of declaring an image under scrutiny to be a cover image when in fact it is stego for a set probability of deciding stego when cover should have been chosen. Given the distributions for cover and stego images, detection theory describes the detector solving this problem. For cover distribution (pdf or PMF) $P_X(\cdot) = P(\cdot|H_0)$ and stego distribution $P_S(\cdot) = P(\cdot|H_1)$ the optimal test is the likelihood ratio test (LRT) [67]:

$$\frac{P_X(Y_1, \ldots, Y_N)}{P_S(Y_1, \ldots, Y_N)} \underset{S}{\overset{X}{\gtrless}} \tau(\alpha)$$

where τ is a threshold chosen to achieve a set false alarm probability, α. In other words, evaluate which hypothesis is more likely given the received data, with a

bias against one hypothesis. Often in practice, a logarithm is taken on the LRT to get the equivalent log likelihood ratio test (LLRT). For convenience we define the log-likelihood statistic:

$$L(Y_1, \ldots, Y_N) \triangleq \log \frac{P_X(Y_1, \ldots, Y_N)}{P_S(Y_1, \ldots, Y_N)} \tag{3.1}$$

and the optimal detector can be written as (with rescaled threshold, τ)

$$\delta(Y_1, \ldots, Y_N) = \begin{cases} H_0 & \text{if } L(Y_1, \ldots, Y_N) > \tau \\ H_1 & \text{if } L(Y_1, \ldots, Y_N) \leq \tau. \end{cases}$$

Applying these results to the steganalysis problem is inherently difficult, as little information is available to the steganalyst in practice. As mentioned before, assumptions are made to obtain a well-posed problem. A typical assumption is that the data samples, $(Y_1, \ldots, Y_N)$, are independent and identically distributed (i.i.d.): $P(Y_1, \ldots, Y_N) = \prod_{n=1}^{N} P(Y_n)$. This simplifying assumption is a natural starting point, commonly found in the literature [10, 63, 21, 75, 46] and is justified in part for data that has been de-correlated, with a DCT transform for example. Additionally this assumption is equivalent to a limit on the complexity of the detector. Specifically the steganalyst need only study histogram based statistics. This is a common approach [35, 69, 21], as the histogram is easy to calculate and the statistics are reliable given the number of samples available in image steganalysis. Therefore in order to develop and apply the detection theory approach, we

assume i.i.d. data throughout this chapter. In general this model is incomplete, and in the next chapter we extend the model to include a level of dependency.

Under the i.i.d. assumption, the random process is completely described by the marginal distribution: the probabilities of a single sample. As we generally consider discrete valued data, our decision statistic comes from the marginal PMF. For convenience we use vector notation, e.g. $\mathbf{y} \triangleq (Y_1, \ldots, Y_N)$, $\mathbf{p}^{(X)}$ with elements $p_i^{(X)} \triangleq \text{Prob}(X = i)$. With this notation the cover and stego distributions are $\mathbf{p}^{(X)}$ and $\mathbf{p}^{(S)}$ respectively.

Let $\mathbf{q}$ be the empirical PMF of the received data, found as a normalized histogram (or type) formed by counting the number of occurrences of different events (e.g. pixel values, DCT values), and dividing by the total number of samples, N. Under the i.i.d. assumption, the log-likelihood ratio statistic is equivalent to the difference in Kullback-Leibler (K-L) divergence between $\mathbf{q}$ and the hypothesis PMFs [18]:

$$L(\mathbf{y}) = N[D(\mathbf{q}\|\mathbf{p}^{(S)}) - D(\mathbf{q}\|\mathbf{p}^{(X)})]$$

where the K-L divergence $D(\cdot\|\cdot)$ (sometimes called relative entropy or information discriminant) between two PMFs is given as

$$D(\mathbf{p}^{(X)}\|\mathbf{p}^{(S)}) = \sum_{i \in \mathcal{Y}} \mathbf{p}_i^{(X)} \log \frac{\mathbf{p}_i^{(X)}}{\mathbf{p}_i^{(S)}}. \tag{3.2}$$

where $\mathcal{Y}$ is the set of all possible events m. We sometimes write $L(\mathbf{q})$ where it is implied that $\mathbf{q}$ is derived from $\mathbf{y}$. Thus the optimal test is to choose the hypothesis with the smallest Kullback-Leibler (K-L) divergence between $\mathbf{q}$ and the hypothesis PMF. So although the K-L divergence is not strictly a metric, it can be thought of as a measure of the "closeness" of histograms in a way compatible with optimal hypothesis testing. In addition to providing an alternative expression to the likelihood ratio test, the error probabilities for an optimal hypothesis test decrease exponentially as the K-L divergence between cover and stego, $D(\mathbf{p}^{(X)}|\mathbf{p}^{(S)})$ increases [6]. In other words, the K-L divergence provides a convenient means of gauging how easy it is to discriminate between cover and stego. Because of this property, Cachin suggested [10] using the K-L divergence as a benchmark of the inherent detectability of a steganographic system. In the i.i.d. context, a data hiding method that results in zero K-L divergence would be undetectable; the steganalyst can do no better than guessing. Achieving zero divergence is a difficult goal (see Chapter 5 for our approach) and common steganographic methods in use today do not achieve it, as we will show. We first demonstrate the detection-theoretic approach to steganalysis by studying a basic but popular data hiding method: the hiding of data in the least significant bit.

3.2 Least Significant Bit Hiding

In this section we apply the detection-theoretic approach to detection of an early data hiding scheme, the least significant bit (LSB) method. LSB data hiding is easy to implement and many software versions are available (e.g. [47, 48, 49, 27]). With this scheme, the message to be hidden simply overwrites the least significant bit of a digitized hiding medium, see Figure 3.1 for an example. The intended receiver decodes the message by reading out the least significant bit. The popularity of this scheme is due to its simplicity and high capacity. Since each pixel can hold a message bit, the maximum rate is 1 bits per pixel (bpp). A disadvantage of LSB hiding, especially in the spatial domain, is its fragility to any common image processing [52], notably compression. Additionally, as we will see, LSB hiding is not safe from detection.

3.2.1 Statistical Model for LSB Hiding

Central to applying hypothesis testing to the problem of detecting LSB hiding is a probabilistic description of the cover and the LSB hiding mechanism. The i.i.d. cover is $\{X_n\}_{n=1}^N$, where the intensity values X_n are represented by 8 bits, that is, $X_n \in \{0, 1, ..., 255\}$. We use the following model for LSB data hiding with

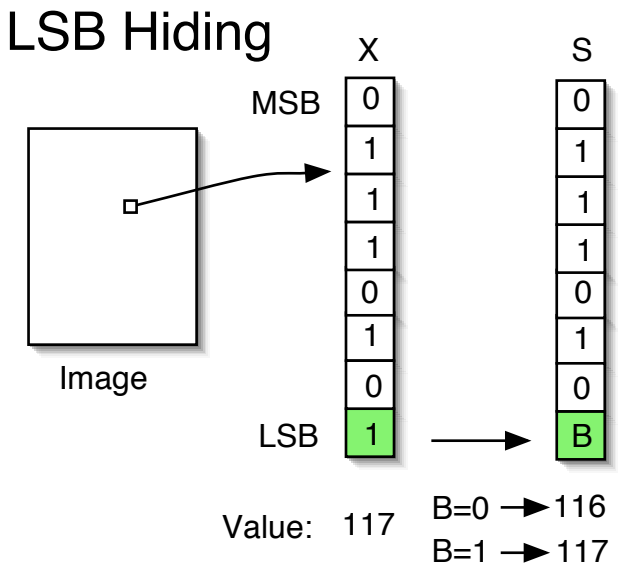

Figure 3.1: Example of LSB hiding in the pixel values of an 8-bit grayscale image.

rate R bits per cover sample. The hidden data $\{B_n\}_{n=1}^{N}$ is i.i.d. and,

$$P_B(b_n) = \begin{cases} R/2 & b_n \in \{0,1\} \\ 1-R & b_n = \text{NULL} \end{cases}$$

With $0 < R \leq 1$. The hider does not hide in cover sample X_n if $B_n = \text{NULL}$, otherwise the hider replaces the LSB of X_n with B_n. With this model for rate R LSB hiding, and again denoting the PMF of X_n as $\mathbf{p}^{(X)}$, then the PMF of the

stego data after LSB hiding at rate R is given by,

$$p_i^{(S_R)} = \begin{cases} \left(1 - \frac{R}{2}\right) p_i^{(X)} + \frac{R}{2} p_{i+1}^{(X)} & i \text{ even} \\ \\ \frac{R}{2} p_{i-1}^{(X)} + \left(1 - \frac{R}{2}\right) p_i^{(X)} & i \text{ odd} \end{cases}$$

For a more concise notation, we can write $\mathbf{p}^{(S_R)} = \mathbf{Q}_R \mathbf{p}^{(X)}$, where $\mathbf{Q}_R$ is a 256×256 matrix corresponding to the above linear transformation.

3.2.2 Optimal Composite Hypothesis Testing for LSB Steganalysis

Since LSB hiding can embed a particularly high volume of data, the steganographer may purposely hide less in order to evade detection; hence we must account for the hiding rate. In this section, for the i.i.d. cover and LSB hiding described above, we extend the hypothesis testing model of Section 3.1 to a composite hypothesis testing problem in which the hiding rate is not known. As with other hiding schemes we consider, we first assume that the cover PMF is known to the detector so as to characterize the *optimal* performance.

Rather than a simple test deciding between cover and stego, we wish to decide between two possibilities: data is hidden at some rate R, where $R_0 \leq R \leq R_1$, or no data is hidden ($R = 0$). The parameters $0 < R_0 \leq R_1 \leq 1$ are specified by the user. We use H_R to represent the hypothesis that data is hidden at rate

R. The steganalysis problem in this notation is to distinguish between H_0 and $K(R_0, R_1) \triangleq \{H_R : R_0 \leq R \leq R_1\}$. The hypothesis that data is hidden is thus *composite* while the hypothesis that nothing is hidden is *simple*. For this case our detector is:

$$\delta(Y_1, ..., Y_N) = \begin{cases} H_0 & \text{if } (Y_1, ..., Y_N) \in A, \\ K(R_0, R_1) & \text{if } (Y_1, ..., Y_N) \in A^c. \end{cases}$$

In [21], Dabeer proves for low-rate hiding that the optimal composite hypothesis is solved by the simple hypothesis testing problem: test H_0 versus H_{R_0}. This greatly simplifies the problem, allowing us to use the likelihood ratio test (or minimum K-L divergence) introduced in Section 3.1.

3.2.3 Asymptotic Performance of Hypothesis Tests

Without having to simulate actual likelihood tests, we can estimate the performance of detection from the K-L divergence. For the case of small divergence, when the hiding introduces little change to the cover statistics, we can employ an asymptotic normality assumption to the decision statistic, $L(\mathbf{y})$. From this we can calculate approximate expressions of the error probabilities for lage N and different R. This allows us to avoid time-consuming Monte Carlo simulations when evaluating detection, especially helpful when comparing between detection methods.

We note the empirical PMF of received data can be written as:

$$\mathbf{q} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{Z}_n$$

where $\mathbf{Z}_n$ is a column vector whose m-th entry is 1 if received data $Y_n = m$ and is zero otherwise. As Y_n are i.i.d., $\mathbf{Z}_n$ are i.i.d., and we have then a sum of i.i.d. random variables. From the central limit theorem, $\mathbf{q}$ converges in distribution to the Gaussian:

$$\mathbf{q} \Longrightarrow \mathcal{N}(E[\mathbf{Z}_1], N^{-1}\mathbf{\Sigma}_R)$$

where $\mathbf{\Sigma}_R$ is the covariance matrix of $\mathbf{Z}_1$. Under hypothesis H_R, $E[\mathbf{Z}_1] = \mathbf{p}^{(\mathbf{S_R})}$, and $\mathbf{\Sigma}_R = \mathrm{diag}(\mathbf{p}^{(\mathbf{S_R})}) - \mathbf{p}^{(\mathbf{S_R})}(\mathbf{p}^{(\mathbf{S_R})})^T$ (for H_0, we note that $\mathbf{p}^{(\mathbf{S_0})} \equiv \mathbf{p}^{(\mathbf{X})}$). Now suppose our decision processing statistic, L, is differentiable at $\mathbf{p}^{(\mathbf{S_R})}$, then [71]

$$L(\mathbf{q}) \Longrightarrow \mathcal{N}(\mu(R), N^{-1}\sigma^2(R))$$

$$\mu(R) = L(\mathbf{p}^{(\mathbf{S_R})})$$

$$\sigma^2(R) = \mathbf{u}_R^T\mathbf{\Sigma}_R\mathbf{u}_R$$

$$\mathbf{u}_R \triangleq \nabla L|_{\mathbf{p}^{(\mathbf{S_R})}}.$$

We are interested in finding expressions for the probabilities of errors for the detector based on $L(\mathbf{q})$. For LSB hiding, we are most concerned with small R, around 0.05. Here the alternative hypotheses are close, and the asymptotic normality result provides good approximations to the error probabilities. For the

large divergence case, in which the error region is several standard deviations from the means, the central limit theorem becomes increasingly inaccurate. Here we can use results from large deviation theory [22] to bound the error probabilities. For example, a commonly used result in hypothesis testing is the Chernoff bound [67], which we use in Section 3.3 for QIM detection. With the Gaussian approximation for large N and small R we have

$$\begin{aligned} P(\text{False Alarm}) &= P(L(\mathbf{q}) < \tau | H_0) \\ &\approx Q\left(\frac{\sqrt{N}(\mu(0) - \tau)}{\sigma(0)}\right) \end{aligned} \tag{3.3}$$

$$\begin{aligned} P(\text{Miss}) &= P(L(\mathbf{q}) \geq \tau | H_R) \\ &\approx Q\left(\frac{\sqrt{N}(\tau - \mu(R))}{\sigma(R)}\right) \end{aligned} \tag{3.4}$$

where $Q()$ is the complementary Gaussian function:

$$Q(u) \triangleq \int_u^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

We now compare the LLRT to the χ^2 (chi-squared) steganalysis test used by Westfeld et al [100] and by Provos' Stegdetect [69]. The decision statistic is

$$L_{\chi^2}(\mathbf{q}) \triangleq \sum_{k=0}^{127} \frac{(q_{2k+1} - q_{2k})^2}{q_{2k} + q_{2k+1}}.$$

If $L_{\chi^2}(\mathbf{q})$ is less than a threshold, then data is declared to be hidden, otherwise the image is considered clean. We know that after hiding at $R = 1$,

$$p_{2k}^{(S_R)} = p_{2k+1}^{(S_R)} = \frac{p_{2k}^{(S_R)} + p_{2k+1}^{(S_R)}}{2}, \quad 0 \leq k \leq 127$$

The χ^2 statistic is a measure of the closeness of the adjacent bins $\{2k, 2k+1\}$; A smaller statistic implies adjacent bin values are nearly equal, and there is a high chance that data is hidden. To compare the two decision statistics, the LLRT, $L(\mathbf{q})$, and χ^2, $L_{\chi^2}(\mathbf{q})$, we use binomial cover PMFs, $B\{255, \theta\}$, $\theta \in (0, 1)$ and mixtures of binomials. We observed that χ^2 performs very close to the optimal LLRT. While the χ^2 statistic does not depend on the cover PMF and for a given cover PMF its performance appears to be close to the LLRT, the choice of threshold required to guarantee a target performance depends on the host PMF as shown by the following example. Suppose our target is to minimize the error sum $P(\text{Miss})+P(\text{False Alarm})$. In Figure 3.2, we plot this error sum as a function of the threshold for the LLRT and χ^2 for binomial B{255,0.5} and B{255,0.8} cover PMFs. For the LLRT, the threshold $\tau = 0$ minimizes the error sum for *any* cover PMF, whereas for χ^2 the minimizing threshold depends on the cover PMF.

To summarize, if the cover PMF is known, then there is little loss in using the suboptimal χ^2 test. In practice this means that if we have good models for the cover PMF and lookup tables for choosing the threshold (depending on the cover PMF), the χ^2 performs close to the LLRT. However, the cover PMF usually

varies substantially over image databases, and hence we are more interested in completely data driven test that attain the target performance; both the statistic and the threshold have to be chosen based on the data to achieve the desired performance. With this in mind, we note two points, which motivate our work in the following section.

1. The likelihood ratio depends on the cover PMF. However the threshold τ can be chosen independent of the cover PMF. For example to minimize the error sum $P(\text{Miss}) + P(\text{False Alarm})$, we can choose $\tau = 0$.

2. The χ^2 statistic does not depend on the cover PMF, but to obtain a target performance, the threshold τ has to be chosen depending on the cover PMF. Thus χ^2 does *not* resolve the problem of not knowing the cover PMF; it simply transfers it to the choice of τ.

3.2.4 Practical Detection Based on LLRT

Given the discussion in Section 3.2.2, we now restrict our attention to the simple hypothesis testing problem: test H_0 versus H_R, $R > 0$. We propose tests based on the estimation of the LLRT statistic and exhibit their superiority over χ^2. We also develop estimates of the hiding rate R.

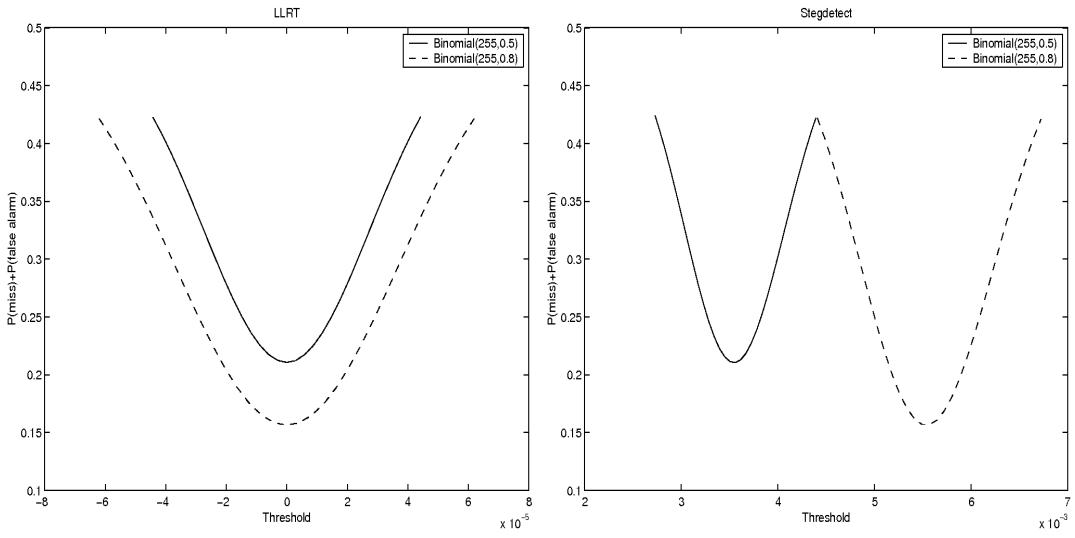

Figure 3.2: Unlike the LLRT, the χ^2 (used in Stegdetect) threshold is sensitive to the cover PMF .

3.2.5 Estimating the LLRT Statistic

The pervading problem with the optimal LLRT test is that we do not know the cover PMF in practice. However, there are two factors that help us to develop good practical tests based on the optimal LLRT.

1. The perturbations introduced by LSB hiding are small compared to the cover PMF, and therefore we can estimate the cover PMF well. We show below that a number of simple estimates of the cover PMF based on the assumption that the cover PMF is 'smooth' work well. We have observed this smoothness for many images, it is however difficult to rigorously justify this assumption.

2. For the optimal LLRT, the threshold that minimizes

$$aP(\text{Miss}) + (1-a)P(\text{False alarm}),\ a \in [0,1]$$

does not depend on the cover. In particular, for $a = 0.5$, the optimal threshold $T = 0$. In contrast, for goodness-of-fit measures such as χ^2, the choice of the threshold depends on a and the cover PMF, and there is no known way of making this choice.

With the above motivation, we propose to form an estimate $\widehat{\mathbf{p}}^{(X)}$ of the cover PMF $\mathbf{p}^{(X)}$ and then use the following estimated version of the decision statistic $L(\mathbf{y})$ as an approximate LLRT statistic:

$$L_{\text{approx.}}(\mathbf{y}) = D(\mathbf{q}\|\widehat{\mathbf{p}}^{(X)}) - D(\mathbf{q}\|\mathbf{Q}_R\widehat{\mathbf{p}}^{(X)}).$$

We consider three possible estimates for $\mathbf{p}^{(X)}$, all of which give good results.

1. For natural images the PMF is usually low pass. On the other hand, random LSB hiding introduces high frequency components in the histogram. Hence one simple estimate $\widehat{\mathbf{p}}^{(X)}$ is to pass the empirical PMF $\mathbf{q}$ through a low pass 2-tap FIR filter with taps $(0.5, 0.5)$. We note that normalization is required after the filtering.

2. Another regularity constraint that we can impose on the cover PMF is that local slope is preserved. That is,

$$\mathbf{p}_{k+3}^{(X)} - \mathbf{p}_{k}^{(X)} = 3(\mathbf{p}_{k+2}^{(X)} - \mathbf{p}_{k+1}^{(X)}),\ k = 0, 4, 8, ..., 252.$$

   This can be concisely written as $\mathbf{A}\mathbf{p}^{(X)} = 0$, where $\mathbf{A}$ is a 64×256 matrix corresponding to the regularity constraint. Under this constraint, a natural estimate of $\mathbf{p}^{(X)}$ is to project $\mathbf{q}$ on to the null space of $\mathbf{A}$. We again need normalization and removal of negative components after this filtering.

3. We also propose a non-linear approach that adapts to the underlying cover PMF. We note that LSB hiding only affects the 8^{th} bit plane. Therefore, we impose the regularity constraint that the cover PMF is such that we can obtain the cover PMF by spline interpolation of the first seven bit planes. The corresponding estimate $\widehat{\mathbf{p}}^{(X)}$ is obtained by subsampling $\mathbf{q}$, then interpolating using splines, and then normalizing.

We refer to all these tests as the approximate LLRT.

Simulation Results

In this section we report and discuss a number of simulation results for four thousand images from a DOQQ image set, as well as digital camera and scanned images.

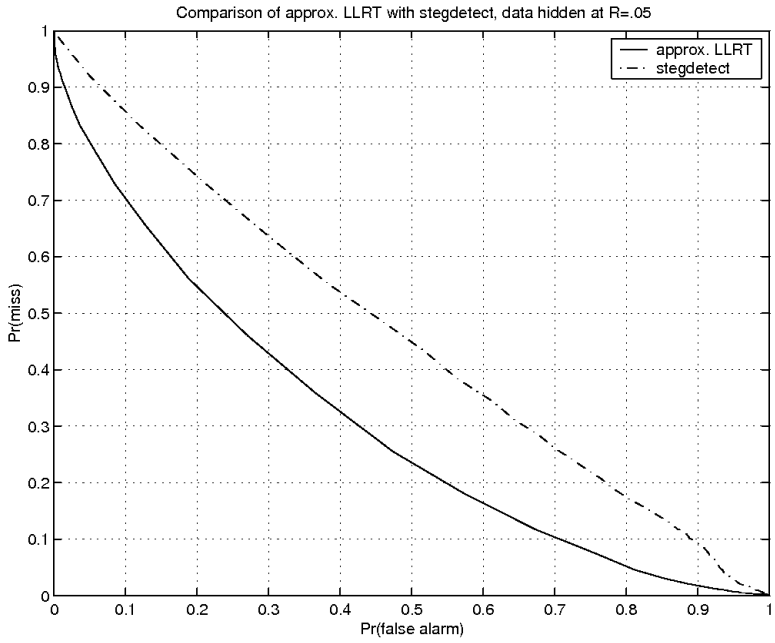

Figure 3.3: Approximate LLRT with half-half filter estimate versus χ^2: for any threshold choice, our approximate LLRT is superior. Each point on the curve represents a fixed threshold.

In Figure 3.3 we compare the approximate LLRT test based on the half-half filter for estimating $\mathbf{p}^{(X)}$ with χ^2. For each point on the curve, the threshold has been fixed over the entire database. In the ROCs, the best detector reaches the origin (0 false alarms and misses), and the worst detector is on the line connecting the upper-left corner to the lower-right. At this rate, and other rates we test, the LLRT outperforms χ^2. For a fixed cover PMF, both these tests perform closely. However, for the database of images we have used, the cover PMF varies substantially from image to image. Thus these simulations suggest that χ^2 is more

sensitive to the choice of the threshold than our approximate LLRT test. This is not surprising since we know that to attain a target performance, the choice of the threshold in LLRT does not depend on the cover PMF. For example, if we choose $T = 0$ for the approximate LLRT in the case when the hiding rate is 0.05, then we found the operating point to be $P(\text{Miss}) = 0.4043$ and $P(\text{False Alarm}) = 0.3219$. From Figure 3.3 we can verify that the tangent to the operating curve at this point is of slope approximately 1 as predicted by the theory. The approximate LLRT is therefore closer to the goal of finding a data driven test.

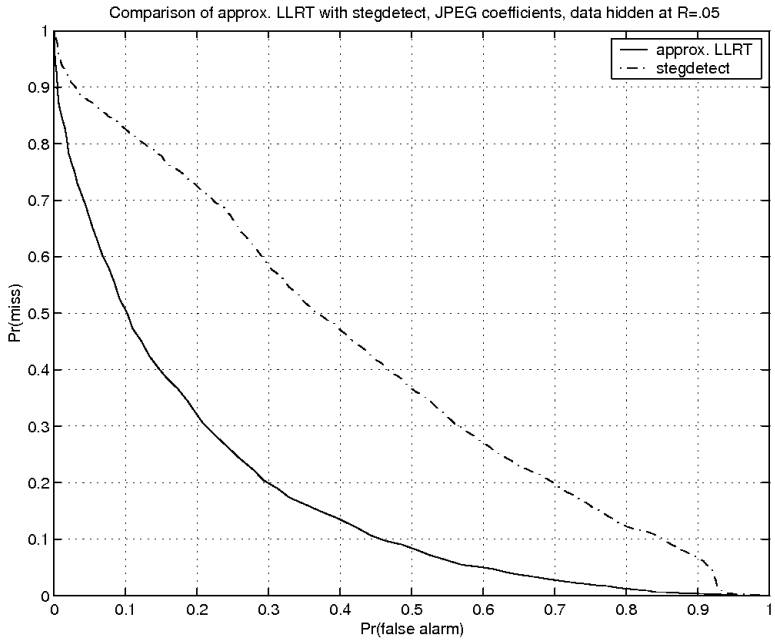

Figure 3.4: Hiding in the LSBs of JPEG coefficients: again LRT based method is superior to χ^2.

Figure 3.4 shows that the story remains unchanged if we hide in the LSB of the JPEG coefficients of images (compressed with quality factor 75).

In principle, instead of the simple hypothesis tests as above, we could use the following generalized LLRT (GLLRT) ([67]) type test:

$$\max_{R\in(0,1]} D(\mathbf{q}|\mathbf{p}^{(X)}) - D(\mathbf{q}|\mathbf{p}^{(S_R)}) \underset{X}{\overset{S}{\gtrless}} \tau. \tag{3.5}$$

This GLLRT performs very close to the (simple) approximate LLRT tests we have developed (which use R_0 instead of R). This is not surprising given our earlier finding that the optimal composite hypothesis testing problem considered in Section 3.2.2 is solved by the simple hypothesis testing problem.

Additionally, we can use the argument R that maximizes (3.5) as an estimate of the actual embedding rate. We find this to work reasonably well in practice, see Figure 3.5.

Different image sets

To increase the diversity of our testing material, we also tested the approximate LLRT on databases of 128 scanned images and 3000 digital camera images. Though the performance of the LRT-based method on these databases is still much better than the χ^2 test, we found the approximate LLRT detection power suffered a drop for these image sets. The change in cover statistics may decrease the divergence between cover and stego. Additionally, our cover estimation as-

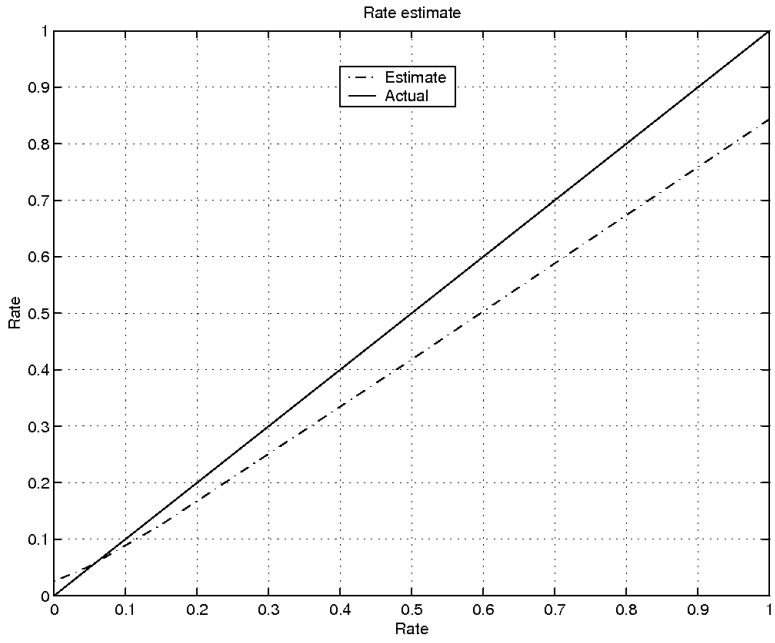

Figure 3.5: The rate that maximizes the LRT statistic (3.5) serves as an estimate of the hiding rate.

sumes a certain smoothness to the cover PMF, that may not be as valid over all possible cover statistics.

Different estimation functions

To gauge the efficacy of various estimation functions, we compare the approximate LLRT scheme based on different estimates of $\mathbf{p}^{(X)}$. The spline estimates of $\mathbf{p}^{(X)}$ and the half-half low pass filter estimates perform nearly identically. We have observed that the local slope preserving filter is slightly worse off. This suggests that there might be little to gain from choosing a different cover estimate.

Comparison with RS

The focus to this point has been to develop the optimal methodology for steganalysis of i.i.d. data. RS analysis [33] is a non-hypothesis testing based algorithm for estimating the length of an LSB hidden message that exploits correlation. By comparing with our approximate LLRT we can gauge the improvement that can be gained by including cover memory. In Figure 3.6, we can see the improvement gained by acknowledging the cover correlation. Other image databases showed similar gains. In Chapter 4 we extend the detection theoretic framework to include a notion of cover memory.

S-tools

We further test our approximate LLRT LSB detection on S-Tools [9], a popular software for data hiding that uses an LSB hiding method. One function of S-tools is to hide data into color images. Though we generally focus on grayscale images, we test our detection on S-tools because it is a freely available and popular tool. We embedded random data at maximum rate using S-tools in a database of hundreds of color images. This database is comprised of images from two sources: images from a Corel photo CD and images scanned from photographs. The Corel photos had to be converted from PCD format to BMP for embedding, and the scanned images are converted from PNG format. The approximate LLRT

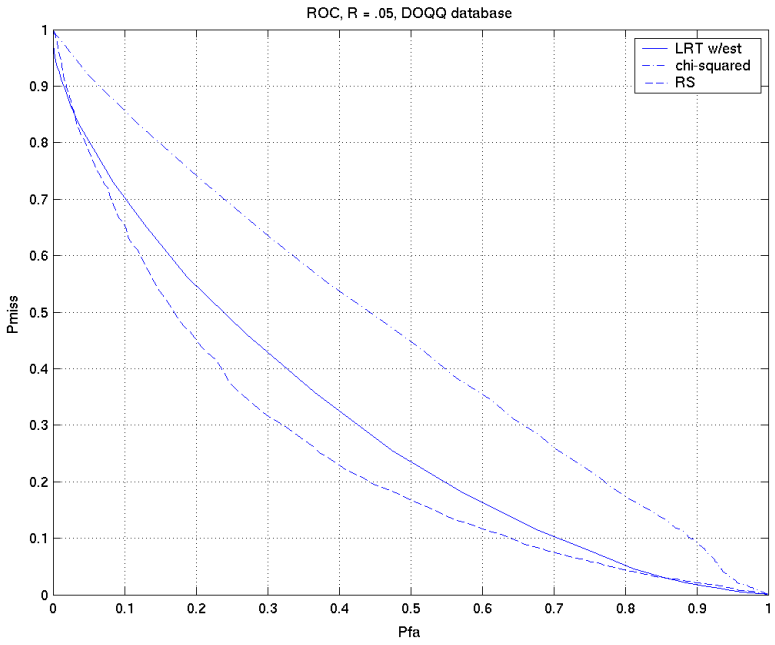

Figure 3.6: Here RS analysis, which uses cover memory, performs slightly better than the approximate LLRT. A hiding rate of 0.05 was used for all test images with hidden data.

is adapted to three color planes by simply adding the LLRT statistics from each color plane. The ROC of the approximate LLRT is in Figure 3.7.

The results are not as good as we expect from our testing on grayscale images, where we find error free detection over hundreds of images with full embedding. It can be seen from the ROC that after a certain threshold, the number of false alarms increases quickly. By inspecting the images that were falsely labeled as stego by our detector, we find that in all cases, the images are those that had been converted from PCD format. Inspecting the histograms of these images, we

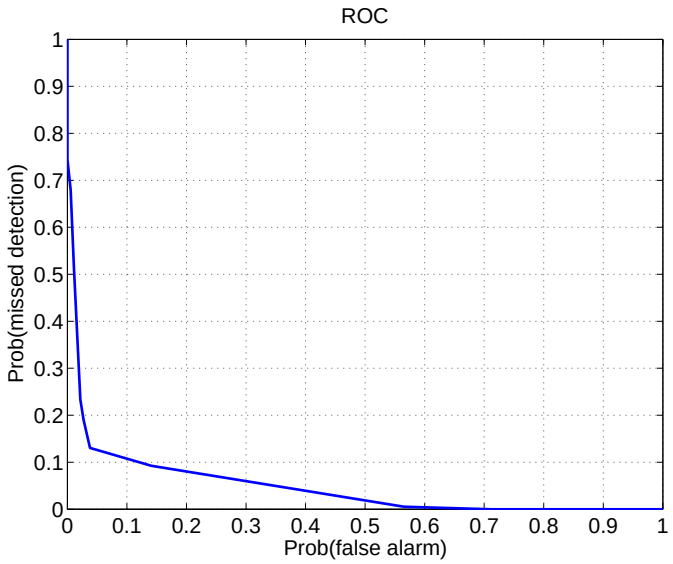

Figure 3.7: Testing on color images embedded at maximum rate with S-tools. Because format conversion on some color images tested on causes histogram artifacts that do not conform to our smoothness assumptions, performance is not as good as our testing on grayscale images.

noticed a distinctive artifact of periodic spikes in the histogram; see Figure 3.8 for an example. Clearly, in this case, our assumption of a smoothly varying histogram does not hold. We continue to find throughout this study many discrepancies in performance from one set of images to another. This presents a great difficulty in evaluating the performance of steganalysis tools. For example, in these tests we could "achieve" perfect performance by simply choosing not to use the Corel image database. Therefore, to fairly evaluate our steganalysis, throughout our testing we attempt to use as diverse a database of images as is feasible.

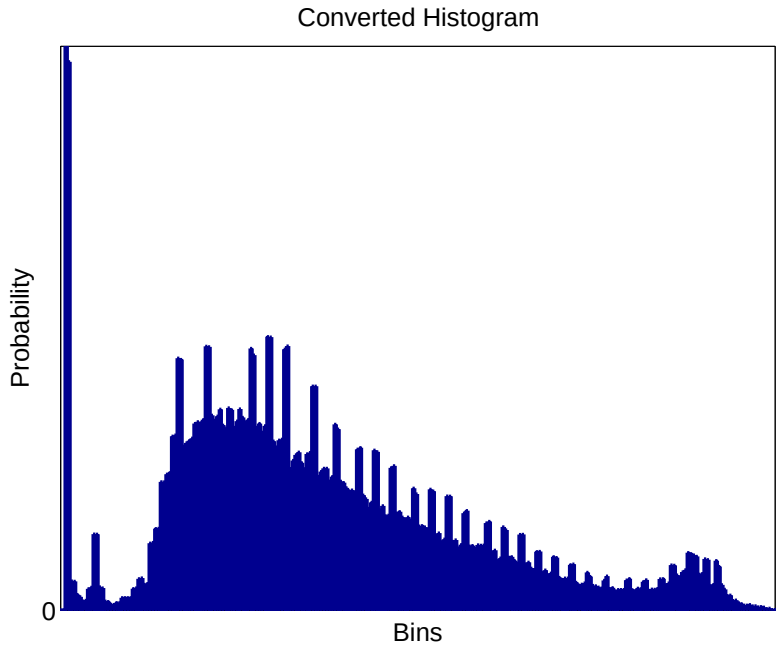

Figure 3.8: Conversion from one data format to another can sometimes cause idiosyncratic signatures, as seen in this example of periodic spikes in the histogram.

3.2.6 LSB Hiding Conclusion

By analyzing the optimal hypothesis test for steganalysis and making justified assumptions about the PMFs of typical real images, we have formulated tools for detecting LSB steganography. Our method performs better than previous hypothesis testing approaches in two ways.

1. They lead to smaller probability of miss for the same probability of false alarm.

2. The choice of the threshold is less sensitive to variations in the cover PMF. Thus for typical hiding rates less than 0.1, the choice of threshold $\tau = 0$ leads to good performance.

3.3 Quantization Index Modulation Hiding

Quantization index modulation (QIM) [14] is another popular hiding method. Though more complex and generally with lower capacity than LSB hiding, QIM is much more robust to noise and attacks. The essential idea of quantization based data hiding is to quantize the cover signal with a quantizer indexed by the message. Again, denote the cover and stego signals as $\{X_n\}_{n=1}^N$ and $\{S_n\}_{n=1}^N$, and let b be the message. We have $S(X, b) = q_b(X)$. The stego signal consists 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 [14], can produce a stego signal covering all of the values of the cover signal. Here the quantizers are shifted according to a changing dither level D

$$S(X, m) = q_b(X + D) - D = q(X + D(b)) - D(b) \tag{3.6}$$

There exist more advanced flavors of QIM, which provide advantages to simpler versions, but these are generally designed for watermarking applications. For example, we note that while distortion compensated QIM (DC-QIM) has performance benefits for watermarking, the statistical effect of DC-QIM is so distinctive [98] that we believe it is not suitable for covert data hiding. Most practical implementations we have seen use either simple QIM, or dither modulation, with uniform scalar quantizers, thus we focus on these.

We note also that typically a steganographer is hiding in data transformed to make it suitable for compression. This is done for two reasons. First, hiding in a compression transform spreads the hidden data over a large region, making it less susceptible to local attacks. Second, the image is likely compressed if it is transmitted, and hiding in the compressed medium allows the steganographer to more easily anticipate the effect of compression. (For a thorough exploration of the validity of these assumptions see [102].)

3.3.1 Statistical Model for QIM Hiding

To apply our detection-theoretic approach, we need a description of the effect of uniform scalar QIM hiding on the cover PMF. For uniform scalar QIM, one bit, b, of the message is embedded into each sample. The quantizer function q_b is

$$q_b(X) = \begin{cases} \text{round}(X/\Delta) & b = 0 \\ \text{round}(X/\Delta) + \Delta/2 & b = 1 \end{cases}$$

see Figure 3.9.

For the steganographer, the choice of Δ represents a trade off between robustness and distortion to the image. Later we show Δ's relationship to the steganographic security. The relationship of Δ to robustness depends on what we define as robust. For robustness to the classical additive white Gaussian noise

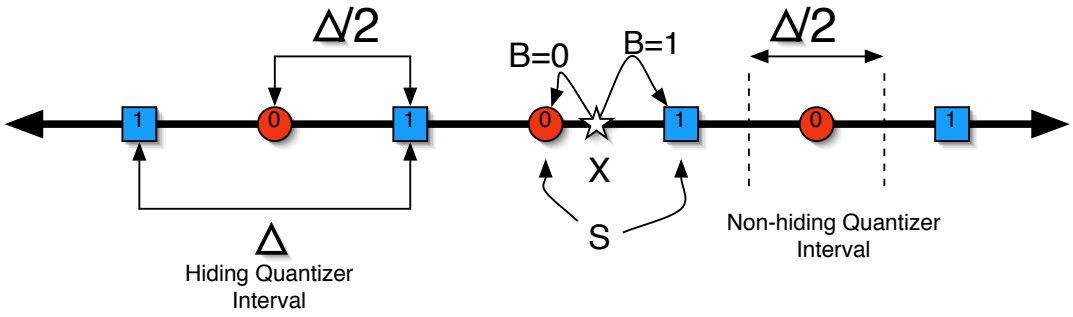

Figure 3.9: Basic scalar QIM hiding. The message is hidden in choice of quantizer. For QIM designed to mimic non-hiding quantization (for compression for example) the quantization interval used for hiding is twice that used for standard quantization. X is cover data, B is the bit to be embedded, S is the resulting stego data, and Δ is the step-size of the QIM quantizers.

(AWGN) channel with power σ^2, the probability of error is:

$$P(\text{error}) \sim Q\left(\sqrt{\frac{(\Delta/2)^2}{4\sigma^2}}\right)$$

[14]. For robustness to channels designed for optimal attack, the robustness also scales with Δ^2 [42]. The point to note is that the steganographer faces a robustness constraint when choosing Δ.

Since QIM hiding is generally used in a transform domain (e.g. DCT, DFT) rather than the spatial domain, the sample space is no longer $[0, 255]$, but instead the real line $\mathbb{R}$, or an interval on the real line. In this case, the distribution of cover and stego is a pdf, rather than a PMF. Practically however, the steganalyst uses a histogram with several bins of width w to estimate the pdf. For our study of QIM steganalysis, we employ PMFs based on a sample space of the bin

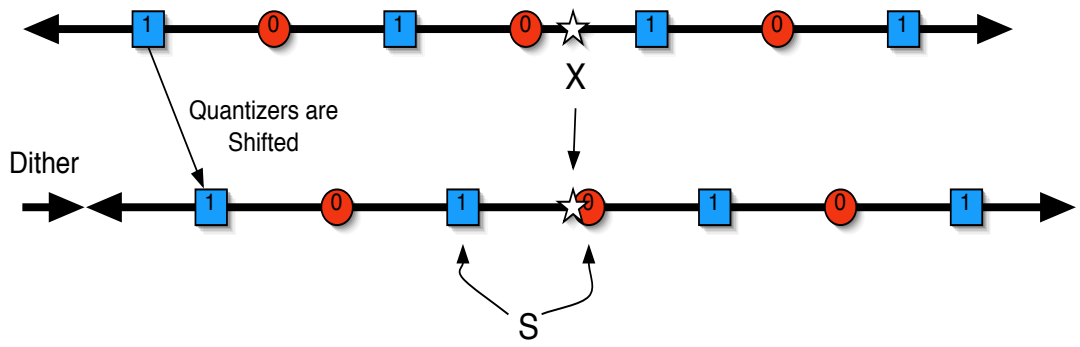

Figure 3.10: Dithering in QIM. The net statistical effect is to fill in the gaps left behind by standard QIM, leaving a distribution similar, though not equal to, the cover distribution.

centers. We explore this approximation in greater detail in Chapter 5; here it is a warranted assumption on the steganalyst. Now the vector notation used for PMFs in our study of LSB hiding is not convenient and we denote the PMF as $P_X(x) \triangleq \text{Prob}(X = x)$.

Again, we are assuming for now that X_n are i.i.d. so the one-dimensional PMF is sufficient for classification. As with LSB hiding, we can find the PMF of S as a function of P_X. We begin by examining a non-hiding, uniform scalar quantizer. The output levels are the integer multiples of the step-size, Δ^*, and the probability of a given output, C, 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

$\mathcal{X}^*(c) \triangleq [c - \Delta^*/2, c + \Delta^*/2)$ then the PMF is

$$P_C(c) = \begin{cases} \sum_{x \in \mathcal{X}^*(c)} P_X(x), & c \in k\Delta^* \\ 0 & \text{else} \end{cases} \tag{3.7}$$

Where k is any integer. If now the choice of quantizer is used to hide binary data, B, we split the original quantizer into two 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

$$P_S(s) = \begin{cases} \frac{1}{2} \sum_{x \in \mathcal{X}(s)} P_X(x) & s \in \frac{k\Delta}{2} \\ 0 & \text{else} \end{cases} \tag{3.8}$$

Where $\mathcal{X}(s) \triangleq [s - \Delta/2, s + \Delta/2)$ is the analogous range for the new Δ. 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 [81] and [54], 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 (PR) variable uniformly distributed over $[-\Delta/4, \Delta/4)$ so that the output covers all the values of the input, and does not leave tell-tale signs of quantization. Under our PMF approximation,

$P_D(d) = 2w/\Delta$. With this dithering, any S is valid. For every received value of S there is one and only one valid value of D and one valid value of B that could have made that s. For any valid S, $P_S(s) = P_B(\text{required}\, b) \cap \sum_{x \in \mathcal{X}(s)} P_X(x) \cap P_D(\text{required}\, d)$. Assuming equiprobable message data $P(b)$ is $\frac{1}{2}$ for either 0 or 1. Similarly since P_D is uniform, the probability is the same for any d so after plugging in we have

$$P_S(s) = \frac{w}{\Delta} \sum_{x \in \mathcal{X}(s)} P_X(x) \tag{3.9}$$

We note D can alternatively be distributed over $[-\Delta/2, \Delta/2)$ with no change on the distribution of S [98]. Additionally in this case it has been shown that S and X are statistically independent [40].

3.3.2 Optimal Detection Performance

Armed with equations (3.7), (3.8), and (3.9) we can find the performance of a detector operating in two scenarios. The first is distinguishing between cover values that have been quantized versus QIM data embedding (without dithering). The second case is distinguishing between an unquantized cover and a cover with dithered QIM data embedded.

As mentioned in Section 3.1, the optimal detector is the log-likelihood ratio test. Before we analyze the performance of this detector for some example PMFs,

we can gain some insight into what is detectable simply by inspecting the log-likelihood ratio (3.1):

Case I Quantized cover versus non-dithered QIM hiding:

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

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

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

Case II Non-quantized cover versus dither modulation hiding:

$$L(\mathbf{y}) = \log \prod_{k=1}^{N} \frac{(w/\Delta) \sum_{y_k - \Delta/2 \leq x < y_k + \Delta/2} P_X(x)}{P_X(y_k)}$$

The product terms are exactly the ratio of the average (over Δ) to the original. Dither modulation hiding therefore acts as a moving average filter on the PMF. Alternatively it is the convolution of a uniform PMF with twice the range of the dither sequence, and is therefore similar to SS hiding. Intuitively, cover PMFs with high frequency components relative to Δ are much easier to detect than a smoothly varying PMF. Indeed, as is noted in [43], a uniformly distributed cover would be impossible to detect.

As mentioned above, QIM hiding is typically done in a transform domain. Compressed data generally has values concentrated towards the mean. That is, the PMF tends to be unimodal with a large spike at the center. See, for example, the histogram of DCT coefficients of an image in Figure 3.11. 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 σ to Δ. As mentioned above, Δ is related to the robustness of the hiding.

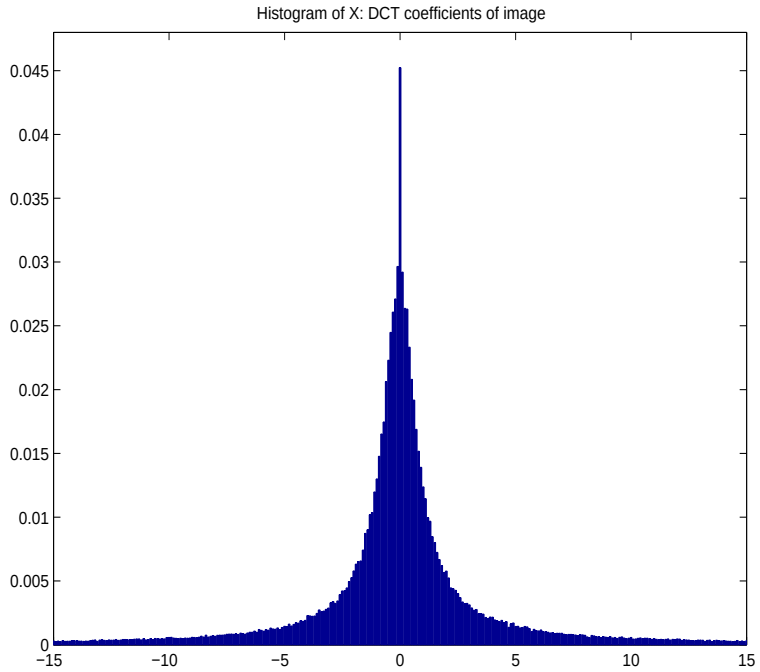

Figure 3.11: The empirical PMF of the DCT values of an image. The PMF looks not unlike a Laplacian, and has a large spike at zero.

To quantify this observation, we can find the performance of the detector for a given cover distribution. As the priors are not known, as a metric we use the sum of the probabilities of false alarm and missed detection. As mentioned in Section 3.2.2 for a known PMF, we find upper bounds on these probabilities by using Chernoff bounds (for details see [67]). Chernoff bounds allow us to find a bound on the performance even at very low probabilities of error, which is not possible with simulations. For a set of images, we use the normalized histogram as an empirical PMF and simulate the detectability over many PR generated "images". We find the detectability is extremely sensitive to the ratio σ/Δ, see Figure 3.12. Here, we are detecting a Laplacian PMF at rate 1. Within a short range of σ/Δ, the detection metric goes from nearly certain detection to almost random detection. Gaussian PMFs have a similar relationship. Because of the low error rates in the low σ/Δ case, the asymptotic normality assumptions used for LSB analysis (Section 3.2.3) would not accurately estimate small σ/Δ. An interesting parallel to note: Mihçak and Venkatesan [61] noticed a similar relationship for *intended* detection of quantization hiding in statistics of discrete wavelet transform (DWT) coefficients subject to Gaussian noise with power σ^2.

Recently Hogan et al [46] also presented results on optimal detection of QIM. For a message to cover ratio (MCR) of -20 dB, they found detector error rates on

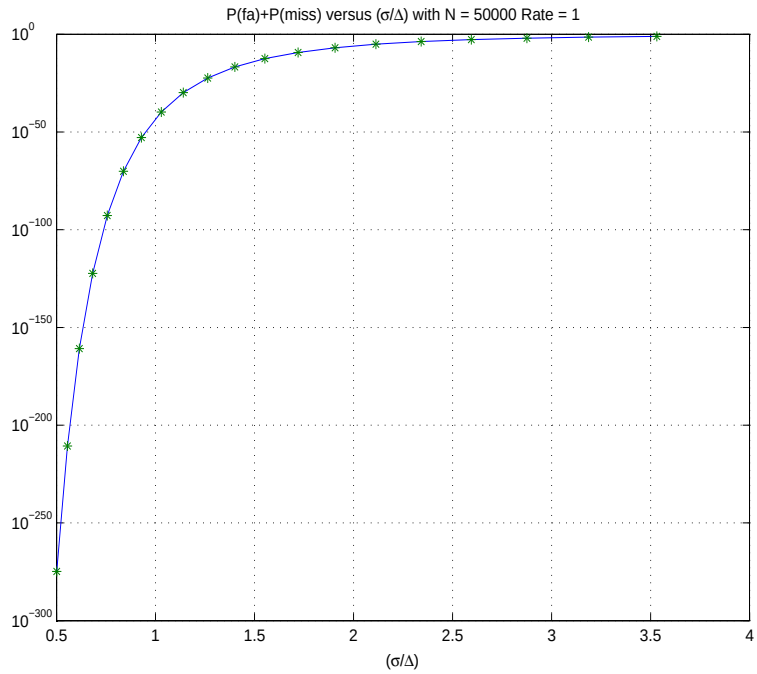

Figure 3.12: The detector is very sensitive to the width of the PMF versus the quantization step-size.

the order of 10^{-1}. The MCR for QIM hiding is:

$$MCR = \frac{\Delta^2}{12\sigma^2}$$

so

$$\frac{\sigma}{\Delta} = \frac{1}{\sqrt{12(10^{-MCR/10})}}$$

and σ/Δ corresponding to MCR of -20 db is 2.9. Though difficult to see from Figure 3.12, these error rates are consistent with our findings.

The hider therefore should choose to embed in either a high variance cover, or use a small Δ. However the choice of covers may be limited, and a smaller Δ

weakens its robustness to external attacks. He or she may choose then to embed less data than is possible in order to avoid detection. Let R, measured in bits per cover sample characterize this. For scalar QIM, $0 < R \leq 1$. As R is reduced, the detectable difference between the hidden statistics and cover statistics is diluted by the cover samples that pass unchanged. We can easily adjust equations (3.8) and (3.9) to reflect this:

$$P_S(s, R) = RP_S(s, 1) + (1 - R)P_X(s) \tag{3.10}$$

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:

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

To estimate error probabilities with the GLRT, we use computer simulation rather than Chernoff bounds, yielding a rougher estimate. We find that 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 cover 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, see Figure 3.13. 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 15,000 bit message in 50,000 cover samples ($R = 0.3$). If the cover is a Gaussian with $(\sigma/\Delta) = 1$ the detector has an error sum (the probability of false alarm plus probability of missed detection) of 0.070. If we hold the number of samples constant but halve the rate to $R = 0.15$, the sum of errors jumps to 0.366, because of the reduction in rate. However this only sends 7,500 bits. To send the entire message the hider has to use 100,000 cover samples. The error sum for the steganalyst taking this into account drops to 0.205.

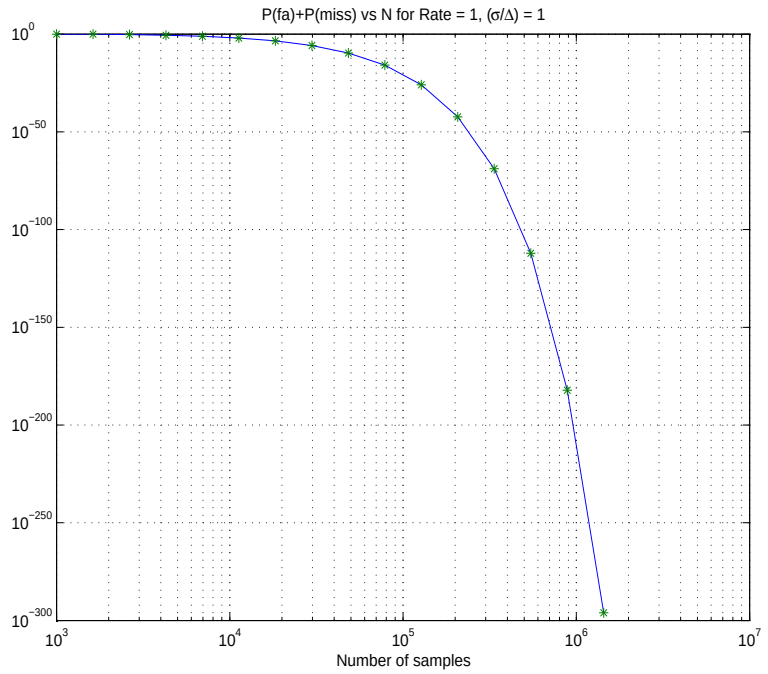

Figure 3.13: Detection error as a function of the number of samples. The cover PMF is a Gaussian with $(\sigma/\Delta) = 1$

3.3.3 Practical Detection

Finally, in implementing these schemes on real world data, certain adjustments must be made to the basic scheme. For example both [81] and [54] exclude low-valued coefficients from embedding. This is done to avoid visual distortion of the final image, and to prevent the characteristic smoothing of probability near the mean we observed above. 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 offers little illumination.

We do not have access to the cover statistics in a practical detection setting. To overcome this, we can attempt to estimate the cover statistics for each received image, or estimate the cover statistics for all images. Some steganalysis is able to estimate the statistics on an image by image basis, as in our work with LSB hiding and the work by Fridrich et al on detecting F5 [35], however there is no general prescription for making such an estimate. As noted in Chapter 2, there has also been some success with classifying between the set of all cover images and all stego images through the use of supervised learning techniques. The idea is to train a machine with several examples of both cover and stego, and the machine learns to discriminate between the two classes. For our experiments we choose to employ supervised learning.

Motivated by the likelihood ratio test, we use an empirical PMF (histogram) as the feature vector. We performed simulated detection of a DCT-based QIM scheme [81]. This scheme is designed to use the same quantization as JPEG. The hiding is designed for a given JPEG quality factor, which translates to a set of quantization step sizes for different DCT frequencies. Higher quality factor implies smaller step size. Since the quality factor used is not known by the steganalyst, Δ is not available. In an attempt to capture the variety of real-world images, we use three databases: digital orthophoto quarter-quadrangle (DOQQ) aerial images used also for testing LSB hiding, Corel PhotoCD (CPCD) images, and images taken with a Canon digital camera (CACD), also used for LSB testing. For a classifier we use Joachim's support vector machine (SVM) implementation [51], SVM^{light}. A linear kernel is used; we found other kernels perform only slightly differently. We perform the following two tests:

With constant design quality factor: In the first experiment, we set the design quality factor at 50 to hide data in images in both the training and testing sets. Since the quality factor is static, Δ is the same for all stego images, aiding the classifier. After hiding, both the cover images and the images with hidden content are 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 3.1 We find that if the design quality factor is constant, the

Final Quality Factor	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 3.1: If the design quality factor is constant (set at 50), a very low detection error can be achieved at all final quality levels. Here '0' means no errors occurred in 500 tests so the error rate is < 0.002

detection with supervised learning gives very low error rates, which remain low even with severe JPEG compression. We understand though, that a constant design quality factor is not usually available for the detector and it is expected to make detection simpler. In the next test, we eliminate this restriction.

With varying design quality factor: We perform this test with several design quality factors. This is achieved by creating training and testing sets by hiding data in images with the design quality factor randomly chosen between 40 and 80. The stego image is then JPEG compressed at a variety of rates between 50 and 100. The same tests are performed as in the known quality factor case. The results are shown in Table 3.2. From this table, we find that if the quantizer step-size varies, 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.

Final Quality Factor	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 3.2: In a more realistic scenario where the design quality factor is unknown, the detection error is higher than if it is known, but still sufficiently low for some applications. Also, the final JPEG compression plays an important role. As compression becomes more severe, the detection becomes less accurate.

3.3.4 QIM Hiding Conclusion

Our detection-theoretic results for i.i.d. covers show that the ease with which QIM can be detected depends strongly on the cover statistics. Specifically, cover 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. For a given PMF type, the inherent detectability is linked to the ratio of the standard deviation to QIM interval: σ/Δ. Though the steganographer can decrease the probability of hidden data being detected by decreasing Δ, robustness is sacrificed. While the assumed knowledge of cover distribution 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 strongly.

3.4 Summary

In this chapter we introduced our detection-theoretic approach to steganalysis, and applied it to the detection of LSB and QIM methods of data hiding. We found both steganographic methods to be theoretically detectable and designed practical schemes with which to detect hiding in real images. Our initial approach to applying detection theory to steganalysis has been to use an i.i.d. model. A natural extension to this is to use a model including at least one level of dependency, an extension we consider in the following chapter.

Chapter 4

Extending Detection-theoretic Steganalysis to Include Memory

As noted in Chapter 2 there are several practical image steganalysis techniques which exploit interpixel correlations in images. Additionally, our work in applying detection theory with an i.i.d. model to detection of LSB and QIM has hinted that extending our model to include a degree of dependency will improve our estimates and methods. In this chapter we employ a Markov model with our detection-theoretic approach.

4.1 Introduction

Existing theoretical benchmarks of optimal detection such as Cachins's ϵ-secure measure [10] model the cover data as i.i.d., and therefore underestimate the attainable steganalysis performance. We now take the logical next step to-

wards computing a more accurate performance benchmark, modeling the cover data as a Markov chain (MC). The Markov model has the advantage of analytical tractability, in that performance benchmarks governing detection performance can be characterized and computed explicitly. In our examples and numerical results, we focus on images as cover data, using a Markov model in which statistical dependency is limited to an adjacent pixel. Clearly, this model does not completely capture interpixel dependencies. However, we find that the performance benchmarks we compute are consistent with the performance of a number of image steganalysis techniques that exploit spatial correlations.

With the extension of analysis to include dependency, we see immediate benefits over current i.i.d. analysis. For example, though spread spectrum (SS) hiding can be detected with reasonable accuracy using current steganalysis techniques [45, 11, 86], security tests derived from i.i.d. analysis determine that spread spectrum hiding is in fact safe from detection. Markov chain analysis, on the other hand, correctly determines SS hiding to be at risk from steganalysis. Additionally, though recently steganographic schemes have been designed to reduce the probability of detection, the efforts have generally focused on matching the one-dimensional histogram [68, 26, 75] or other specific steganalysis statistics [76, 29], which provides no guarantee that a future steganalysis scheme will not be able to detect the hiding by using a different statistic. Furthermore, though the γ_D-

security, proposed by Chandramouli et al [12], does not use an i.i.d. assumption, the measure is with respect to a given detector, and does not gauge the performance of other detectors. On the other hand our analysis predicts optimal steganalysis based on a Markov chain assumption. Although the MC model does not completely characterize image statistics, practical constraints on the ability of the steganalyst to estimate more complex statistical models have limited and may continue to limit the complexity of detectors.

4.2 Detection Theory and Statistically Dependent Data

We here outline our steganalysis approach using a Markov chain model, and show how it relates to common steganalysis techniques.

4.2.1 Detection-theoretic Divergence Measure for Markov Chains

To include interpixel dependency in our analysis, we employ a Markov chain (MC) [16] model of image data. A Markov chain is a random sequence indexed by n, subject to the following condition: $P(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) = P(Y_n|Y_{n-1})$. Under this model, the probability of a given pixel is dependent on an immediately

adjacent pixel. We have used this model to analyze and detect spread spectrum hiding [86]. In independent work, Sidorov has performed MC and Markov random field analysis on detecting LSB hiding [78]. There are a number of reasons to use a MC model. First, the MC model accounts for dependency, yet is very general and flexible. Second, while a MC model is more complex than an i.i.d. model, it is the least complex model incorporating dependencies. Though many have used Markov random fields [70] to model images accounting for a larger neighborhood of dependency than one adjacent pixel, for the steganalyst there is a practical drawback to increasing the levels of dependency. As the model complexity increases, the number of samples required to make an accurate estimate of the statistics also increases. However the number of received samples depends on the image size, and can not be increased by the steganalyst. Thus although the complexity for the steganalyst increases quickly, the benefit does not. For more on the difficulties of multivariate density estimation see [96]. The MC model on the other hand is simple enough to make realistic statistical estimates. This is analogous to a nth order DPCM coding system, in which the benefit of an increase in n, the number of pixels used for prediction, has been shown to quickly drop after 2 or 3 [44]. Finally, a divergence metric which measures the performance of optimal detection, analogous to the K-L divergence for i.i.d. sources, exists [64] for Markov chains, which we examine below.

We first clarify our notation. Let $\{Y_n\}_{n=1}^N$ be a Markov chain on the finite set $\mathcal{Y}$. In our context Y_n are the n-indexed set of pixels obtained by a row or column scanning and $\mathcal{Y}$ are all possible gray scale values (e.g. for an 8-bit image, $\mathcal{Y} = \{0, 1, \ldots, 255\}$). A Markov chain source is defined by a transition matrix, $\mathbf{T}_{ij} \triangleq P(Y_n = i | Y_{n-1} = j)$, and marginal probabilities $p_i \triangleq P(Y_n = i)$. For a realization $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$, let $\eta_{ij}(\mathbf{y})$ be the number of transitions from value i to value j in $\mathbf{y}$. The empirical matrix is $\mathbf{M}(\mathbf{y}) \triangleq (\eta_{ij}(\mathbf{y})/(N-1))$. That is, the i, j-th element represents the proportion of spatially adjacent pixel pairs with a grayscale value of i followed by j, and therefore provides an estimate of the probability $P(Y_n = i, Y_{n-1} = j)$. The empirical matrix thus provides an estimate of the transition matrix and marginal probabilities: $\mathbf{T}_{ij} = P(Y_n = i, Y_{n-1} = j)/P(Y_{n-1} = j)$; $P(Y_{n-1} = j) = P(Y_n = i) = \sum_j P(Y_n = i, Y_{n-1} = j)$. The empirical matrix, similar to the co-occurrence matrix (see citations in [93]), can be recognized as a matrix form of the two-dimensional normalized histogram (or type) used to estimate the joint PMF of an arbitrary source. Intuitively, for sources that are strongly correlated such as pixels, we expect the probability of two adjacent samples having equal, or nearly equal, value to be high. Therefore in the empirical matrix we expect the mass to be more concentrated near the main diagonal (all elements such that $i = j$) in a correlated source then we would expect for an i.i.d. source; see the examples in Figure 4.1.

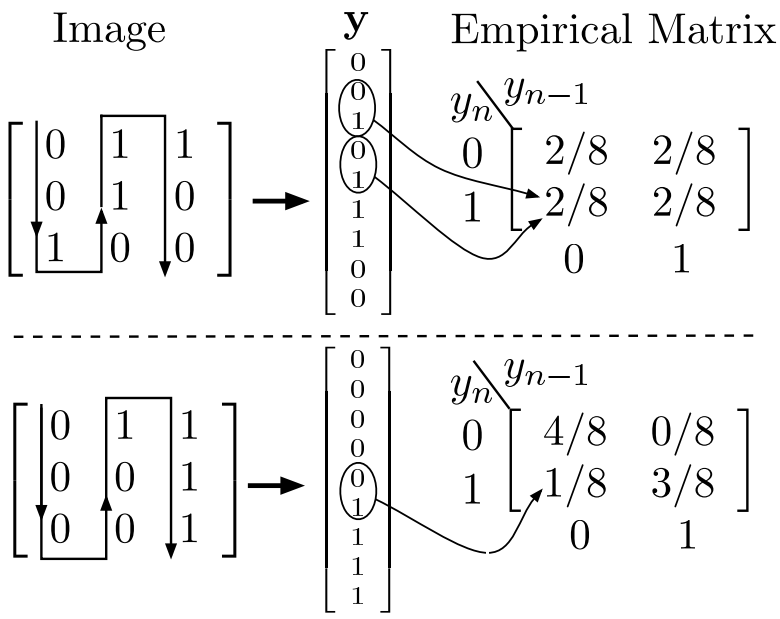

Figure 4.1: An illustrative example of empirical matrices, here we have two binary (i.e. $\mathcal{Y} = \{0, 1\}$) 3×3 images. From each image a vector is created by scanning, and an empirical matrix is computed. The top image has no obvious interpixel dependence, reflected in a uniform empirical matrix. The second image has dependency between pixels, as seen in the homogenous regions and so its empirical matrix has probability concentrated along the main diagonal. Though the method of scanning (horizontal, vertical, zig-zag) has a large effect on the empirical matrix in this contrived example, we find the effect of the scanning method on real images to be small.

The divergence measure we employ to quantify the statistical change introduced by steganography is essentially a distance between the empirical matrices $\mathbf{M}^{(X)}$ and $\mathbf{M}^{(S)}$ of the two hypotheses, cover and stego:

$$D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)}) = \sum_{i,j \in \mathcal{Y}} \mathbf{M}_{ij}^{(X)} \log \left(\frac{\mathbf{M}_{ij}^{(X)}}{\sum_j \mathbf{M}_{ij}^{(X)}} \frac{\sum_j \mathbf{M}_{ij}^{(S)}}{\mathbf{M}_{ij}^{(S)}} \right) \tag{4.1}$$

This divergence has many useful properties for the study of steganalysis in sources with memory, from the point of view of both the steganographer and the steganalyst.

For a constant false alarm rate, the minimal achievable missed detection rate approaches $e^{-ND(\mathbf{M}^{(X)},\mathbf{M}^{(S)})}$ as N, the number of samples, goes to infinity [64, 1], just as in the i.i.d. case with K-L divergence. In other words, under the assumption of a Markov chain model, the performance of the best possible steganalysis is exponentially bounded by this measure.

It can be seen then that $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$ provides a measure to the steganographer of the inherent detectability of a steganographic scheme, given an assumption on the complexity of the detector. This can be seen as an extension of Cachin's [10] ϵ-secure steganography to Markov chains. The gain in using the Markov chain model is the addition of dependency. In other words, if the detector in fact uses dependency, an i.i.d. ϵ-secure hiding scheme overestimates the secrecy of hiding. To prevent this problem, Chandramouli et al [12] suggest the $\gamma_{\mathcal{D}}$ metric. Here the measure of detectability of a steganography method is a bound on the allowed probabilities of false alarm and missed detection for a given detector $\mathcal{D}$. While this certainly avoids the problem of underestimating the power of detectors employing dependency, it is only valid with respect to a given detector. If a different detector is employed, or invented, the security is unknown. The

matrix divergence however bounds the false alarm and missed detection probabilities of the *best possible* detector using one level of dependency. In practice, the steganographer can choose a scheme that minimizes the divergence for a given cover joint distribution model (e.g. Gaussian, Laplacian). Alternatively, given a scheme, the steganographer can choose to use only images that exhibit a small divergence after hiding.

For the steganalyst, $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$ measures the amount of information gained for each additional sample received, just as with the K-L divergence for independent samples. The detector can use this to decide if there is enough gain to justify using a more complex detector. We note that $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$ is equal to the K-L divergence if the samples are indeed independent ($\mathbf{M}_{ij} = p_i p_j$):

$$\begin{aligned}
\sum_{i,j\in\mathcal{Y}} \mathbf{M}_{ij}^{(X)} \log\left(\frac{\mathbf{M}_{ij}^{(X)}}{\sum_j \mathbf{M}_{ij}^{(X)}} \frac{\sum_j \mathbf{M}_{ij}^{(S)}}{\mathbf{M}_{ij}^{(S)}}\right) &= \sum_{i,j} \mathbf{M}_{ij}^{(X)} \log\left(\frac{\mathbf{M}_{ij}^{(X)}}{p_i^{(X)}} \frac{p_i^{(S)}}{\mathbf{M}_{ij}^{(S)}}\right) \\
&= \sum_{i,j} p_i^{(X)} p_j^{(X)} \log\left(\frac{p_i^{(X)} p_j^{(X)}}{p_i^{(X)}} \frac{p_i^{(S)}}{p_i^{(S)} p_j^{(S)}}\right) \\
&= \sum_i p_i^{(X)} \sum_j p_j^{(X)} \log \frac{p_j^{(X)}}{p_j^{(S)}} \\
&= \sum_j p_j^{(X)} \log \frac{p_j^{(X)}}{p_j^{(S)}}.
\end{aligned}$$

Let κ be the ratio of the matrix divergence measure to the K-L divergence. κ represents the gain of employing the more complex model. For example, to achieve the same detector power (i.e. same probabilities of miss and false alarm) requires

κ times as many samples if the detector uses an i.i.d. cover model versus a MC model. In the case of independent data, κ is one, and there is no gain to a detector using statistics beyond a one-dimensional histogram.

4.2.2 Relation to Existing Steganalysis Methods

As mentioned above, using the Markov chain model is analogous to assuming a complexity constraint on the detector. Since dependency is limited to one adjacent pixel, empirical matrices provide sufficient statistics for optimal detection. However even two dimensional joint statistics are difficult to use practically. For practical applications, it is useful to use a subset, or function, of the empirical matrix. Often these subsets or functions are chosen to match a specific hiding scheme, and if done correctly do not sacrifice much detection power. However they certainly cannot improve detection. We now show that many ongoing efforts in steganalysis use such a subset or function.

Many steganalysis schemes and analysis [21, 69, 46, 35, 84] use a histogram, or estimate of the one-dimensional PMF, to discriminate between cover and stego. A one-dimensional histogram is simply the row sums of the empirical matrix: $P(i) = \sum_j \mathbf{M}_{ij}$.

To capture the effect of hiding on interpixel dependencies, some [97, 106] have used difference histograms, that is, instead of a histogram of sample values, a

histogram of the difference of values between samples. As pixels are strongly correlated, the difference between pixels is small, and the histogram is concentrated towards zero. Typically hiding disrupts this concentration, and with appropriate calibration, the hiding can be detected. The difference histogram is the sums of the diagonals of the empirical matrix. That is, the difference histogram is $P(x) = \sum_i \sum_{i-j=x} \mathbf{M}_{ij}$. The concentration at zero in the difference histogram corresponds to the concentration along the main diagonal of the two-dimensional histogram.

To detect LSB hiding, the RS scheme [33] and related sample-pair analysis [24] also use counts of differences between pixel values. Though sample-pair analysis is not limited to adjacent pixels, the authors note the estimate is improved in practice for spatially adjacent samples. In [72], Roue et al use the empirical matrix directly to improve the effectiveness of sample-pair analysis.

Also for LSB detection, Sidorov, explicitly using a Markov chain model [78, 79], uses an entropy-like measure based on ratios of values near the main diagonal of the empirical matrix:

$$-2 \sum_{i=2} \mathbf{M}_{i,i-1} \log \frac{\mathbf{M}_{i,i-1} + \mathbf{M}_{i,i+1}}{2\mathbf{M}_{i,i-1}} + \mathbf{M}_{i,i+1} \log \frac{\mathbf{M}_{i,i-1} + \mathbf{M}_{i,i+1}}{2\mathbf{M}_{i,i+1}}.$$

In [34], Fridrich et al use a calibrated blockiness measure to detect Outguess 0.2 [68]. This blockiness measure is the expected value of of the absolute difference of border pairs, and can be re-written in terms of the empirical matrix generated

from pixels straddling 8×8 block boundaries:

$$Bl = \sum_{x=0,1,\ldots} x \left(\sum_{i} \sum_{|i-j|=x} \mathbf{M}_{ij} \right).$$

In [5], Avcibas et al use image quality metrics to measure the effect of hiding. Though these metrics are not easily related to the empirical matrix, it is notable that the metrics are evaluated between the image under scrutiny, and a low-pass filtered version of the image. To generate the filtered image, each pixel is replaced with a weighted sum of a 3×3 neighborhood surrounding the pixel. In other words, it is assumed that the measurable difference between a given image and the same image with artificially enhanced interpixel dependencies is different for stego images and cover images.

We have argued that analysis using a Markov chain model provides meaningful results under the condition of a steganalyst incorporating one level of dependency for detection. We have seen here that many existing detection methods indeed implicitly employ such a model.

Due to the lack of information available to the steganalyst, practical detection is inevitably suboptimal. However we still expect some relationship between the calculated divergence and the efficacy of state-of-the-art steganalysis. In the following section, we examine the divergence measure two existing steganographic schemes, and compare with current detection methods to test this assumption.

Additionally we compare the calculated divergence under an assumption of independence (Eqn. (3.2)) to the divergence assuming dependency (Eqn. (4.1)) to evaluate the value to the steganalyst in incorporating a more complex statistical model.

4.3 Spread Spectrum

As noted in Chapter 2, spread spectrum data hiding (SS) [19] is an established embedding method, often used for watermarking, but also applicable for steganography [60]. We here measure and study the statistical effect of hiding on the empirical matrix, and relate this to detection experiments we performed.

4.3.1 Measuring Detectability of Hiding

In spread spectrum data hiding, the message data modulates a noise sequence to create a message bearing signal, which is then added to the cover data. Since its introduction, many variants of SS have been proposed, typically in the context of watermarking. The major goal in watermarking is robustness to malicious attacks, rather than statistical invisibility. We therefore focus on three basic models of hiding suggested by Cox et al, shown here for reference. Let $\{D_k, k \geq 0\}$ be a zero mean, unit variance, Gaussian message bearing signal, and $\{X_k, \geq 0\}$ be the

cover samples. Three methods of generating stego data S_k:

$$S_k = X_k + \alpha D_k \tag{4.2a}$$

$$S_k = X_k + (\alpha X_k) D_k = X_k(1 + \alpha D_k) \tag{4.2b}$$

$$S_k = X_k e^{\alpha D_k} \tag{4.2c}$$

where α is a scaling parameter used to adjust the hiding power. This adjustment allows the data hider to adapt the hiding to the cover in order to control the perceptual distortion, the robustness of the message, and security from steganalysis. We have not seen the third method used in practice and, as Cox et al point out, for small αD_i (which we would expect in data hiding) (4.2b) and (4.2c) are effectively similar. We therefore concentrate on the first two methods. In the first method, the adaptation is done globally, i.e. a constant hiding power is used for all cover samples, we refer to this as globally adaptive hiding. In the second method, the hiding power adapts to each cover sample, so we characterize this as locally adaptive. We also note that often the cover image is transformed before data is hidden; for example Costa et al use a whole image discrete cosine transform (DCT). We measure the divergence of four variants of spread spectrum hiding: local and globally adaptive hiding in both the spatial and DCT domains. We have seen globally adaptive spatial SS hiding by Marvel et al in spread spectrum image steganography (SSIS) [60] and more recently by Fridrich et al in a variant

of stochastic modulation [31]. The latter allows for a higher number of bits to be successfully decoded and the capacity is a function of the message signal power. The experiments presented by Cox et al [19] are locally adaptive DCT hiding.

For each variant we calculated the divergence over a range of message signal power. For globally adaptive hiding we hold the message to cover power ratio (MCR) constant. In the locally adaptive case this is not possible so we hold the scale factor α constant. The MCR varies from image to image; we record the average value with the data.

To generalize our approach we would like to simplify the divergence measurement, by eliminating the need to derive the stego empirical matrix. Instead of using statistical analysis of the hiding scheme to generate an expected stego empirical matrix, Monte Carlo simulations of data hiding in several images may provide an accurate means of estimating divergence. To do this, several synthetic images are generated from the empirical matrix of a cover image. Data is hidden in these synthetic images, and stego empirical matrices are calculated from the resulting images. The average divergence between these empirical matrices and the original cover matrix represent an estimate of the divergence introduced by hiding

As mentioned in Section 4.2.1, the divergence measure provides to the steganographer a benchmark of the inherent detectability of hiding. Additionally, it

allows the steganalyst to compare the information gained from each new sample by exploiting dependency, to the information gained using only first order statistics. We present both divergence measures: between empirical matrices (Eqn. (4.1)) and between marginal histograms (Eqn. (3.2)), and the average ratio of these two to show the gain by using dependency at the detector. These measurments are summarized in Tables 4.1 and 4.2. From stochastic modulation [31] (a variant of globally adaptive SS hiding) we have a means of relating message signal power to the capacity. The average hiding rates for MCRs -23, -20, and -17 are 0.91, 0.94, and 0.96 bits per pixel (bpp) respectively.

Globally adaptive, Spatial			
MCR	-23	-20	-17
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	30.20	36.82	43.43
Mean $D(p^{(X)}\|\|p^{(S)})$	2.48	3.27	4.10
Mean ratio	24.23	22.24	20.57
Globally adaptive, DCT			
MCR	-23	-20	-17
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	30.38	36.98	43.46
Mean $D(p^{(X)}\|\|p^{(S)})$	2.49	3.26	4.06
Mean ratio	24.45	22.24	20.56

Table 4.1: Divergence measurements of spread spectrum hiding (all values are multiplied by 100). As expected, the effect of transform and spatial hiding is similar. There is a clear gain here for the detector to use dependency. A factor of 20 means the detector can use 95% less samples to achieve the same detection rates.

From this data we can see many trends. Not surprisingly, the divergence measure always increases with the (MCR); the more powerful a message (and

Locally adaptive, Spatial			
α	0.375	0.05	0.1
Mean MCR	-22.74	-20.33	-14.63
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	27.92	32.17	42.88
Mean $D(p^{(X)}\|\|p^{(S)})$	1.48	1.93	3.34
Mean ratio	33.49	30.31	22.45
Locally adaptive, DCT			
α	0.375	0.05	0.1
Mean MCR	-28.52	-26.13	-20.19
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	5.61	5.87	6.52
Mean $D(p^{(X)}\|\|p^{(S)})$	1.39	1.78	3.41
Mean ratio	13.10	10.21	5.90

Table 4.2: For SS locally adaptive hiding, the calculated divergence is related to the cover medium, with DCT hiding being much lower. Additionally the detector gain is less for DCT hiding.

subsequently a higher hiding rate), the more obvious the hiding becomes. Additionally, though the measured divergence introduced by globally adaptive is roughly the same for both spatial hiding and transform hiding, locally adaptive divergence changes depending on the hiding domain. Locally adaptive spatial hiding is slightly less divergent than globally adaptive hiding (for similar MCR) however locally adaptive DCT is much less. We expect from these divergence measurements that detection is more difficult for locally adaptive hiding, particularly DCT, than for the other cases. Finally in all cases there is an advantage to including dependencies in detection. In the best case, about 95% fewer samples can be used to achieve the same performance. Even in locally adaptive DCT hiding, where the advantage is least, a gain of 5.9 means only about a sixth of the

samples are required. Below we analyze the underlying statistical changes caused by hiding in order to explain these findings.

4.3.2 Statistical Model for Spread Spectrum Hiding

Globally adaptive hiding is analogous to inserting zero mean additive white Gaussian noise (AWGN) with power α^2. The net statistical effect is a convolution of the message signal distribution, $\mathcal{N}(0, \alpha^2)$, with the cover distribution [45]. Deriving the exact empirical matrix of the stego signal is complicated somewhat by the necessity of quantization and clipping as a final step, to return to the same sample space, $\mathcal{Y}$, as the source. For example, in hiding in pixels, the stego values must be rounded to integers between 0 and 255. When necessary to prevent ambiguity, we delineate the unquantized stego signal as S'. The probability density function (pdf) of the stego signal before quantization is:

$$
\begin{aligned}
f_{S'}(s'_1, s'_2) &= \int\int \mathrm{Sum}(t_1, t_2) dt_1 dt_2, \\
\mathrm{Sum}(t_1, t_2) &= \sum_k \sum_l \mathbf{M}^{(X)}_{kl} \delta(t_1 - k, t_2 - l) \frac{1}{2\pi\alpha^2} e^{-\left\{ \frac{(s'_1 - t_1)^2 + (s'_2 - t_2)^2}{2\alpha^2} \right\}} \\
f_{S'}(s'_1, s'_2) &= \frac{1}{2\pi\alpha^2} \sum_k \sum_l \mathbf{M}^{(X)}_{kl} \exp - \left\{ \frac{(s'_1 - k)^2 + (s'_2 - l)^2}{2\alpha^2} \right\}.
\end{aligned}
$$

In other words, there is a white joint Gaussian pdf centered at each point in the cover empirical matrix, and scaled by the empirical matrix value. This can be seen as a blurring of the cover empirical matrix. This is directly analogous to

spatial lowpass filtering of images by convolution with a Gaussian function [41, Sec. 4.3.2]. After rounding to pixel values, the empirical matrix of the stego signal is:

$$\mathbf{M}_{ij}^{(S)} = \int_{i-1/2}^{i+1/2} \int_{j-1/2}^{j+1/2} f_{S'}(s_1', s_2') ds_1' ds_2'. \tag{4.3}$$

Since the message signal is white, that is, uncorrelated, its empirical matrix is spread evenly; there is no greater probability for values near the main diagonal. Hiding weakens the dependencies between the cover samples, which causes a spreading from the main diagonal of the empirical matrix, as seen in Figure 4.2.

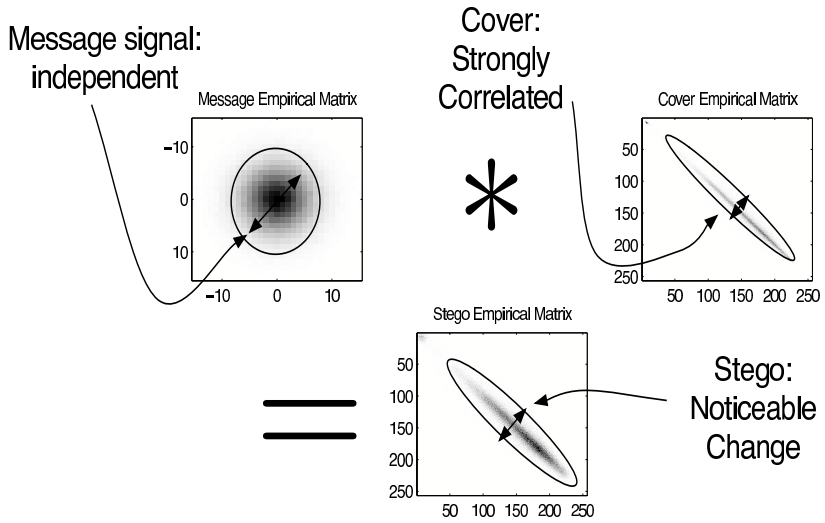

Figure 4.2: Empirical matrices of SS globally adaptive hiding. The convolution of a white Gaussian empirical matrix (bell-shaped) with an image empirical matrix (concentrated at the main diagonal) results in a new stego matrix less concentrated along the main diagonal. In other words, the hiding weakens dependencies.

Locally adaptive hiding can also be viewed as zero mean AWGN, however it is non-stationary, since the noise power $(\alpha X_k)^2$ depends on X_k. Instead we view it as multiplicative noise, with mean of one. Let $B_k \sim \mathcal{N}(1, \alpha^2)$ be a multiplicative message signal, then $S_k = X_k B_k$, The cumulative distribution function (cdf) of (pre-quantized) S is

$$F_{S'_1,S'_2}(s'_1, s'_2) =$$
$$\int_0^\infty \int_0^\infty \left[\sum_{i=0}^{\min(\lfloor s'_1/b_1 \rfloor, 255)} \sum_{j=0}^{\min(\lfloor s'_2/b_2 \rfloor, 255)} \mathbf{M}_{ij}^{(X)} \right] \exp - \left\{ \frac{(b_1-1)^2 + (b_2-1)^2}{2\alpha^2} \right\} db_1 db_1$$

the pdf is

$$f_{S'_1,S'_2}(s'_1, s'_2) = \frac{\partial^2 F_{S'_1,S'_2}(s'_1, s'_2)}{\partial s'_1 \partial s'_2}$$

and the empirical matrix of the quantized S can be found with (4.3). To simplify the expressions, we have assumed α is such that the probability of b_1, b_2 at values less than zero are negligible. This assumption is warranted by the typical α values used in hiding, which are chosen small enough to avoid visual distortion. For a given cover empirical matrix $\mathbf{M}^{(X)}$ these expressions can be evaluated numerically.

From the equations, the statistical effect is not transparent, however as seen in Figure 4.3 hiding still blurs the cover matrix, shifting probability away from the main diagonal. However the effect is less strong.

We can now summarize the statistical effect of SS hiding, and relate this to our findings in Section 4.3.1. In a general sense, SS data hiding adds an

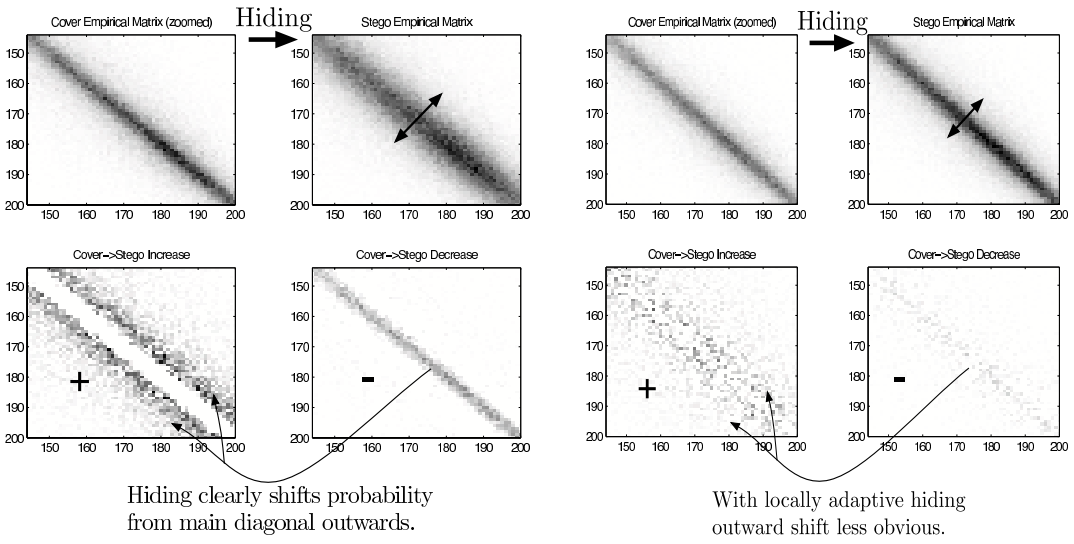

Figure 4.3: Global (left) and local (right) hiding both have similar effects, a weakening of dependencies as seen as a shift out from the main diagonal. However the effect is more pronounced with globally adaptive hiding.

independent and identically distributed (i.i.d.) message signal to a non-i.i.d. cover. It is not surprising then that the statistical effect is a decrease in the dependence of the cover. For globally adaptive hiding this effect is very clearly seen in a shift of probability away from the main diagonal. For locally adaptive hiding, the adaptation causes the additive message sequence to become dependent on the cover. Effectively the message sequence is de-whitened, that is, correlations are introduced, and the effect is weakened. This can be seen to explain the smaller divergence measured for locally adaptive hiding compared to global.

For the linear transformations typically used, such as DCT and discrete Fourier transform (DFT), the addition of a Gaussian message signal in the transform

domain is equivalent to adding a Gaussian message signal in the spatial domain. So globally adaptive hiding in DCT coefficients statistically has the same effect as hiding in pixels. However, locally adaptive hiding is effectively a multiplicative Gaussian signal, which is not equivalent in both domains. This helps explain why the calculated divergence was nearly equal for globally adaptive hiding in either domain, but differed greatly for locally adaptive hiding.

Finally we found the most noticeable effect of spread spectrum hiding is a spreading from the main diagonal of the empirical matrix. Since the histogram is just the collection of sums of each row of the empirical matrix, this effect is missed by studying only marginal statistics. That is, the spreading along each row is not visible when the row is summed into a single point. This explains the gain of using dependency in detection.

4.3.3 Practical Detection

We now compare the measurements of the optimal detectability of SS hiding to experiments using a practical detector. We find the practical experiments follow the estimates above. We also compare experiments for a detector using dependencies with a simpler detector to judge the expected gain in detection.

To achieve optimal detection of data hiding, the detection-theoretic prescription is to calculate the empirical matrix of a suspected image and calculate the

divergence between this and the empirical matrices of both the cover and the stego. Whichever is "closer", i.e. has a smaller divergence measurement, is the optimal decision. From the analysis above we can evaluate the stego empirical matrix given the cover matrix. The cover statistics however are not known in a practical scenario. As with QIM detection in Section 3.3.3 we turn to supervised learning to overcome our lack of knowledge.

For the experiments we need an image database, a learning algorithm, and a feature vector to train the machine. In the image database, we want to represent the vast variety of real images as well as possible. We expand our previous image set to a mix of four separate sources:

1. digital camera images, partitioned into smaller sub-images
2. scanned photographs
3. scanned, downsampled, and cropped photographs
4. images from the Corel volume Scenic Sites

All images are converted losslessly to PNG format and color images are converted to grayscale. The entire database is approximately 1400 images. Half of these are used for training and half for testing. Within both the training and testing sets, half are cover images and half are (distinct) stego images.

For a classifier we use the same SVM implementation, SVM^{light} with linear kernel, as we did for QIM (Section 3.3.3).

Since the optimal hypothesis test finds the minimum divergence between PMF estimates, we are motivated to use PMF estimates to train the SVM. For our experiments with a detector not using dependency, we can use the appropriate PMF estimate: the normalized histogram of pixel values, a 256-dimensional feature vector. Unfortunately for the detector using dependency, the empirical matrix is too large (256^2 dimensions) to use directly. As with the other steganalysis schemes mentioned in Section 4.2.2, we use a reduced version of the empirical matrix for a classification statistic. We have noted that image empirical matrices are very concentrated toward the main diagonal, and that hiding tends to spread the density away from this line. To capture this effect, the feature vector should then include the region immediately surrounding the main diagonal.

To generate the empirical matrix, we need a method of generating a one-dimensional chain from an image, i.e. a scan. We first use a vertical scanning, as in Figure 4.1, for the experiments. We recognize that images have anisotropic dependencies not captured by vertical scanning, so we also explore different feature vectors that combine horizontal, vertical, and diagonal pairs, in order to more accurately characterize pixel dependencies.

For an empirical matrix $\mathbf{M}$ calculated from an image, first the 6 highest probabilities on the main diagonal ($\mathbf{M}_{ii}$) are chosen, and for each of these the following 10 nearest differences are picked:

$$\{\mathbf{M}_{i,i},\, \mathbf{M}_{i,i-1},\, \mathbf{M}_{i,i-2}, \ldots, \mathbf{M}_{i,i-10}\}$$

All together this gives a 66-dimensional vector. We wish to also capture changes along the center line. To do this we subsample the remaining main diagonal values by four:

$$\{\mathbf{M}_{1,1},\, \mathbf{M}_{5,5},\, \mathbf{M}_{9,9}, \ldots, \mathbf{M}_{253,253}\}$$

see Figure 4.4. The resulting total feature vector is 129-dimensional, a manageable size that still captures much of the hiding effect. A comparison of the feature vectors used to evaluate the performance of both detectors, using and not using dependencies, is shown in Figure 4.5. In addition to generating an empirical matrix based on adjacent pixels, we experimented with an empirical matrix generated from a pixel and an average of its four nearest neighbors. This is done in an attempt to capture a possible gain to using a more complex model, while still falling into our framework.

We tested the same four SS variants as in the previous sections. To relate these experiments to other work done, we based our hiding power on that reported in the literature. Spread spectrum image steganography (SSIS) [60] is an implemen-

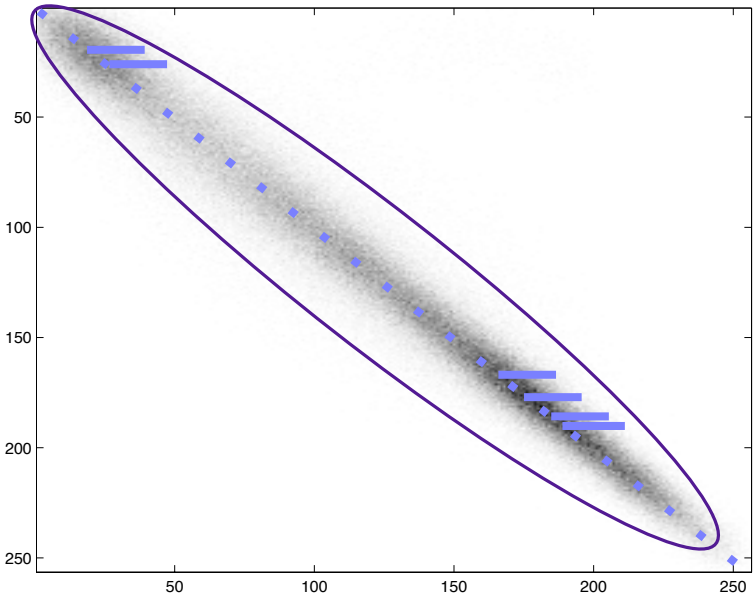

Figure 4.4: An example of the feature vector extraction from an empirical matrix (not to scale). Most of the probability is concentrated in the circled region. Six row segments are taken at high probabilities along the main diagonal and the main diagonal itself is subsampled.

tation of globally adaptive hiding. In the experiments presented by Marvel et al, the MCR reported is always greater than -23 dB, so we choose this as a worst case. For the locally adaptive DCT scheme, we look to the experiments of Cox et al, [19], and use $\alpha = 0.1$, which gives an MCR of roughly -21 dB. In the spatial domain, we choose α to achieve a similar MCR.

Our results are summarized in the receiver operating characteristics (ROC) curves in Figure 4.6 for the detector based on the empirical matrix and the histogram, respectively.

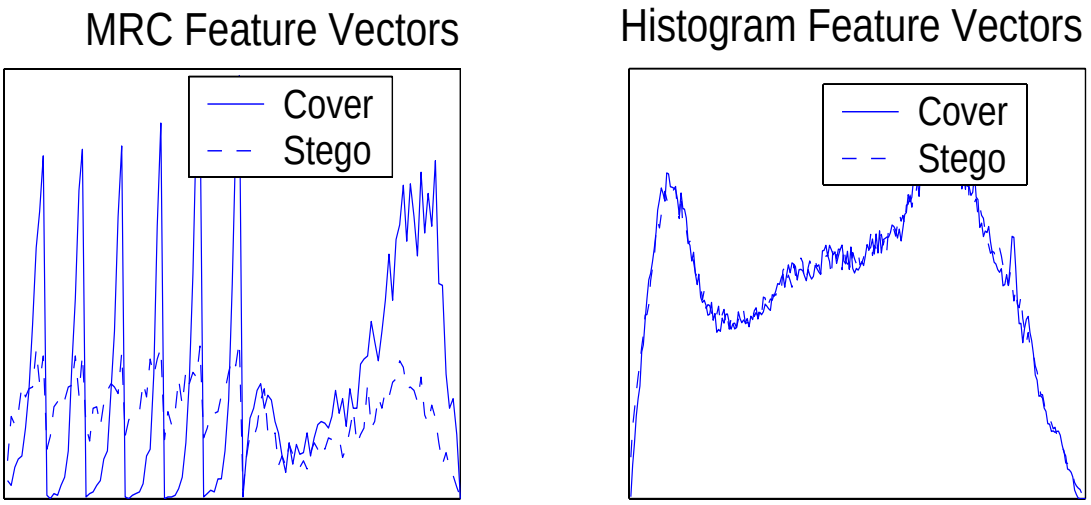

Figure 4.5: The feature vector on the left is derived from the empirical matrix and captures the changes to interdependencies caused by SS data hiding. The feature vector on the right is the normalized histogram and only captures changes to first order statistics, which are negligible.

Anisotropic Dependencies

Since the vertical scan method will not capture all directions of image dependency, we explore different features that incorporate different aspects of dependency. We first look at generating the same feature vector as in the above experiments, but instead scanning the image into a vector using horizontal, or zig-zag scanning (as is done for DCT coefficients in JPEG detection [94]).

In Figure 4.7 we compare the ROCs of three detectors based on vertical, horizontal, and zigzag scans on locally adaptive transform hiding, the hiding scenario with weakest detector performance. All methods perform approximately

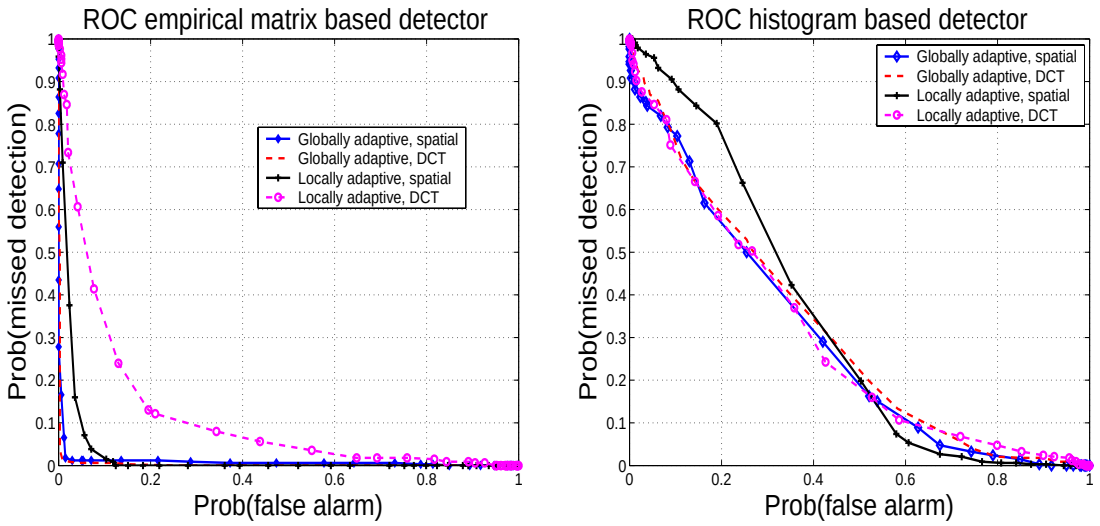

Figure 4.6: ROCs of SS detectors based on empirical matrices (left) and one-dimensional histograms (right). In all cases detection is much better for the detector including dependency. For this detector (left), the globally adaptive schemes can be seen to be more easily detected than locally adaptive schemes. Additionally, spatial and DCT hiding rates are nearly identical for globally adaptive hiding, but differ greatly for locally adaptive hiding. In all cases detection is better than random guessing. The globally adaptive schemes achieve best error rates of about 2-3% for P(false alarm) and P(miss).

the same, with horizontal scan being slightly better than the other two. We find this same trend for locally adaptive spatial hiding as well as global hiding in either domain.

Though there seems to be little difference between directional dependency individually, it may be possible to improve performance by combining different directional information. We therefore also look into methods combining different scans, to see if information from one scan is complementary to another. We looked at three new feature vectors designed to combine statistics based on horizontal,

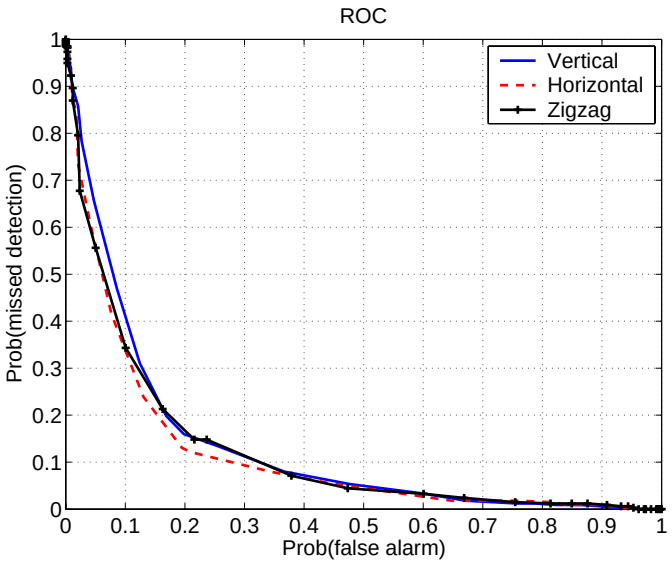

Figure 4.7: Detecting locally adaptive DCT hiding with three different supervised learning detectors. The feature vectors are derived from empirical matrices calculated from three separate scanning methods: vertical, horizontal, and zigzag. All perform roughly the same.

vertical, and diagonal dependencies. The first feature vector is a simple concatenation of the three feature vectors used above based on vertical, horizontal, and zigzag scanning. We denote this as the concatenation feature vector. We point out that this feature vector is three times longer than the standard feature vector, making accurate estimates difficult. The next feature vector uses an empirical matrix formed as the average empirical matrix of all three scans; we refer to this as the averaging feature vector. Since only the derivation of the empirical matrix is different from the standard method, this feature will be the same length. The last feature vector is formed by choosing the 4 highest probabilities on the main

diagonal of each of the three individual empirical matrices, and choosing the 10 nearest differences. Recall that the standard feature vector chooses the 10 nearest differences of the top 6 probabilities, and additionally includes a sub-sampling of the main diagonal. So this final feature vector takes sections we believe to contain the most information from each of the three different scanning vectors. This final feature vector, denoted as cut-and-paste feature vector, is 132-dimensional, nearly the same as the standard 129-dimensional vector. In Figure 4.8 we compare the detector performance on locally adaptive hiding in the spatial and transform domains. For hiding in DCT coefficients, all the detectors perform roughly the same. The three feature vectors combining vertical, horizontal, and zigzag scans perform nominally worse than the standard vector based on horizontal scanning alone. So, although more information is accounted for, there is no gain in using these combined feature vectors (more on this presently). We find these same relative performance characteristics for globally adaptive spatial and DCT hiding. For locally adaptive spatial hiding however, we found the cut-and-paste feature vector to perform much worse than the others. The cut-and-paste feature vector does not include as much information from the main diagonal of the empirical matrix. It can be inferred then that changes to the main diagonal are important to detection of locally adaptive spatial hiding.

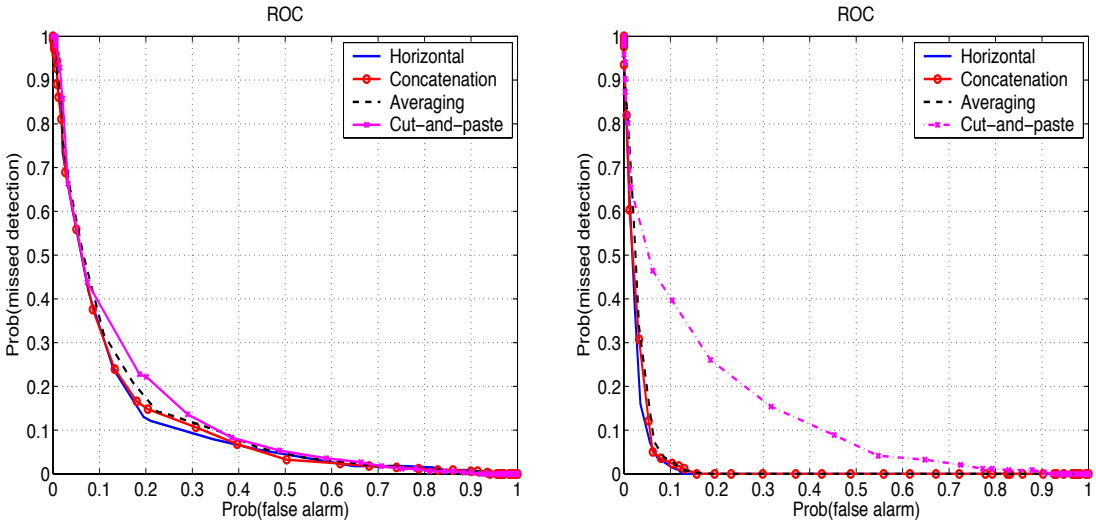

Figure 4.8: ROCs for locally adaptive hiding in the transform domain (left) and spatial domain (right). All detectors based on combined features perform about the same for transform domain hiding. For spatial domain hiding, the cut-and-paste performs much worse.

In addition to combining features, we also look at combining the soft detector output. The SVM classifier gives each image under scrutiny a single number determining its likelihood of being cover or stego. A large negative number indicates likely cover, a large positive number is strongly believed to be stego. By simply adding these outputs from two different classifiers detection may be improved. In other words, the two classifiers will reinforce their results if both have large outputs with the same sign. If the detectors disagree however, the numbers will offset. In particular, this may help in the default case of using a threshold of zero to decide between cover and stego. In experiments performed, we find that generally there is no benefit. However by adding the soft decision statistic from

classifiers based on the horizontal scan feature vector and the cut-and-paste feature vector, better zero threshold performance is acquired for detection of globally adaptive schemes than any other classifier we use, see Table 4.3

Hiding Method	Horizontal		Cut-and-paste		Summed	
	P(f.a.)	P(miss)	P(f.a.)	P(miss)	P(f.a.)	P(miss)
Global adapt., spatial	0.050	0.012	0.050	0.053	0.036	0.012
Global adapt., DCT	0.041	0.006	0.050	0.003	0.011	0

Table 4.3: A comparison of the classifier performance based on comparing three different soft decision statistics to a zero threshold: the output of a classifier using a feature vector derived from horizontal image scanning; the output of a classifier using the cut-and-paste feature vector described above, and the sum of these two. In this particular case, adding the soft classifier output before comparing to zero threshold achieves better detection than either individual case.

Finally, we compare the results of the detector based on the empirical matrix based on one adjacent pixel, and that generated from an average of four adjacent pixels. In Figure 4.9, we compare the results of both detectors for locally adaptive DCT hiding; the results for the other three variants are similar. The detectors perform closely, suggesting the simple Markov chain model captures the important changes introduced by hiding.

Comparison

We note that our results for detecting spatial globally adaptive hiding, error rates on the order of 1 to 5%, are similar to those of Harmsen and Pearlman in [45] detecting SSIS in color images. For detection they used a statistic based

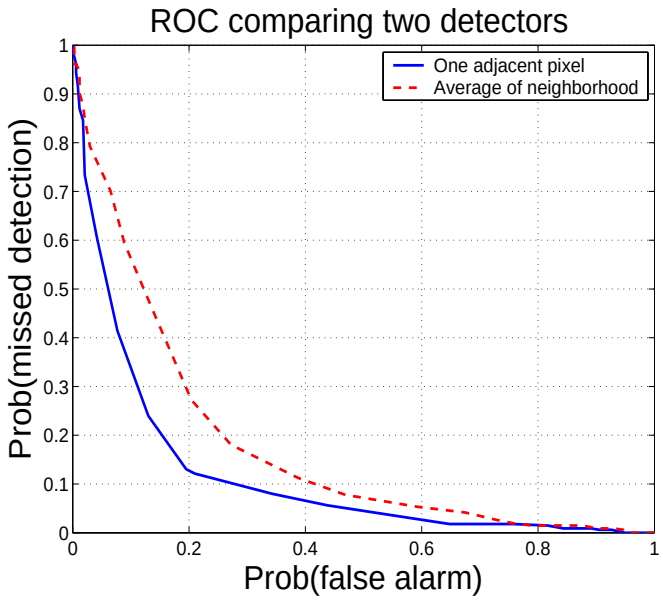

Figure 4.9: A comparison of detectors for locally adaptive DCT spread spectrum hiding. The two empirical matrix detectors, one using one adjacent pixel and the other using an average of a neighborhood around each pixel, perform similarly.

on color plane statistics. Though the detection tests are not directly analogous since our tests are strictly on grayscale images, it is likely that a similar effect to the weakening of dependencies between pixels happens between color planes. Celik et al [11] perform detection of stochastic modulation, statistically the same as spatial globally adaptive hiding. Stochastic modulation allows a greater embedding capacity for a smaller MCR (or larger peak signal to noise ratio PSNR). As such Celik et al tested with a lower MCR, so although their detection error rates (P(false alarm)= 0.17, P(miss)=0.12) are higher than ours, it is difficult to directly compare.

4.3.4 SS Hiding Conclusion

We find the results of a practical detector matches that which our divergence measurements and analysis led us to expect. For the steganographer it may first seem that locally adaptive DCT hiding is the superior choice for hiding. However there are two important points to mention: First, unlike globally adaptive hiding, locally adaptive hiding only meets a target MCR on average. The MCR for each image varies and this may make detection more difficult. Second, although the globally adaptive hiding rate is a function of the message signal power, the locally adaptive hiding rate is not readily available, and may in fact be less than globally adaptive hiding for a given MCR. In all cases there is clearly a gain to the steganalyst to using a model of dependency for detection. In the following section we perform a similar analysis to a hiding scheme specifically designed to evade detection.

4.4 JPEG Perturbation Quantization

Recently Fridrich et al [38] introduced an implementation of their perturbation quantization hiding method that creates stego images that mimic a double-compressed clean image. We measure the divergence of this method and show these are related to practical detection results presented by Kharrazi et al [53]. As

with spread spectrum hiding, we study the statistical effect of hiding to explain these findings.

4.4.1 Measuring Detectability of Hiding

As the name implies, perturbation quantization is a variant of quantization index modulation (QIM) [14]. Standard QIM hiding in JPEG images has a distinctive statistical effect, and we have shown in Section 3.3 that it can be detected. Double compressed PQ however is specifically contrived to minimize the statistical difference between the stego image, and an image that has simply been compressed twice. This is achieved by embedding in coefficients that ideally have the same distribution after a second compression as they do by data hiding.

In [53], Kharrazi et al measure the detection rates for three blind methods of steganalysis used on a variety of steganography schemes. The term "cover" is somewhat ambiguous for PQ JPEG hiding. The original source, from which the stego image is generated, is a once compressed image. However PQ is designed to mimic twice compressed images, which the authors argue occur naturally [38]. Because of this ambiguity, Kharrazi et al measure the detection rates of two cases: comparing with the source (single-compressed) images, and comparing with re-compressed (i.e. double-compressed) images. For the first case, detection is found to be possible, but by no means certain. For example, in one case the sum of

errors (false alarm and missed detection) is about 0.3. For the second case, the detection rates are essentially random. In other words, guessing or flipping a coin is just as effective for steganalysis. For details, please see their paper [53]. We note that the detection schemes are blind to the method, and one would expect better results from a scheme specifically designed to detect PQ JPEG. However these results provide an idea of the practical detectability of this scheme. As with SS above (Section 4.3.1), we measure the divergence introduced by PQ JPEG hiding. In Table 4.4, we summarize the results. Q_1 and Q_2 are the JPEG quality levels used for the first and second compressions. Both these cases correspond to a large number of embeddable coefficients, and all available coefficients are used. For the 75, 50 trial, the average embedding rate is 0.11 bits per pixel (bpp), 0.38 bits per non-zero DCT coefficient (bpnz-DCT). In the second trial (88, 76) the average rate is 0.13 bpp and 0.35 bpnz-DCT.

Single-compressed cover		
Q_1, Q_2	75,50	88,76
Mean MCR	-15.63	-17.89
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	14.64	13.18
Mean $D(p^{(X)}\|\|p^{(S)})$	4.66	3.03
Mean ratio	4.23	6.66

Double-compressed cover		
Q_1, Q_2	75,50	88,76
Mean MCR	-19.26	-21.43
Mean $D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$	3.04	3.89
Mean $D(p^{(X)}\|\|p^{(S)})$	0.63	0.63
Mean ratio	5.28	7.46

Table 4.4: Divergence measures of PQ hiding (all values are multiplied by 100). Not surprisingly, the divergence is greater comparing to a twice compressed cover than a single compressed cover, matching the findings of Kharrazi et al. The divergence measures on the right (comparing to a double-compressed cover) are about half that of the locally adaptive DCT SS case in which detection was difficult, helping to explain the poor detection results.

As in the spread spectrum case, we found that the measure of theoretically optimal detection of data hiding in Markov random chains corresponds to experiments in the non-idealized case. This again suggests that the model is a useful tool in judging the inherent detectability of a steganographic method. Additionally there is a gain for a steganalyst to use dependency for detection, up to 7.5 times gain in this example. We now explore how this low divergence is obtained.

4.4.2 Statistical Model for Double JPEG Compressed PQ

As mentioned above, the source for data hiding is an image that has undergone JPEG compression. During JPEG compression, the image is broken into small blocks, each of which undergoes a 2-d discrete cosine transform (DCT). These DCT coefficients are then quantized to reduce the number of bits used to store or transmit the image (for details see [94]). An inverse DCT of these coefficients reproduces a spatial domain image. However the spatial domain (pixel) values are no longer integers, due to the quantization in the DCT domain. To display or otherwise use the image in the spatial domain, the pixel values are rounded to the nearest integer in the bit depth range (e.g. $\{0, 1, \ldots, 255\}$ for an 8-bit image). Now the DCT coefficients (of the pixel image) are no longer exactly quantized, but instead are randomly spread around the quantized value. So in summary, if an image is compressed, then decompressed, the DCT values are randomly

distributed around their quantized values as seen in Figure 4.10. Asymptotically, this density is a white Gaussian centered at the quantized value [38].

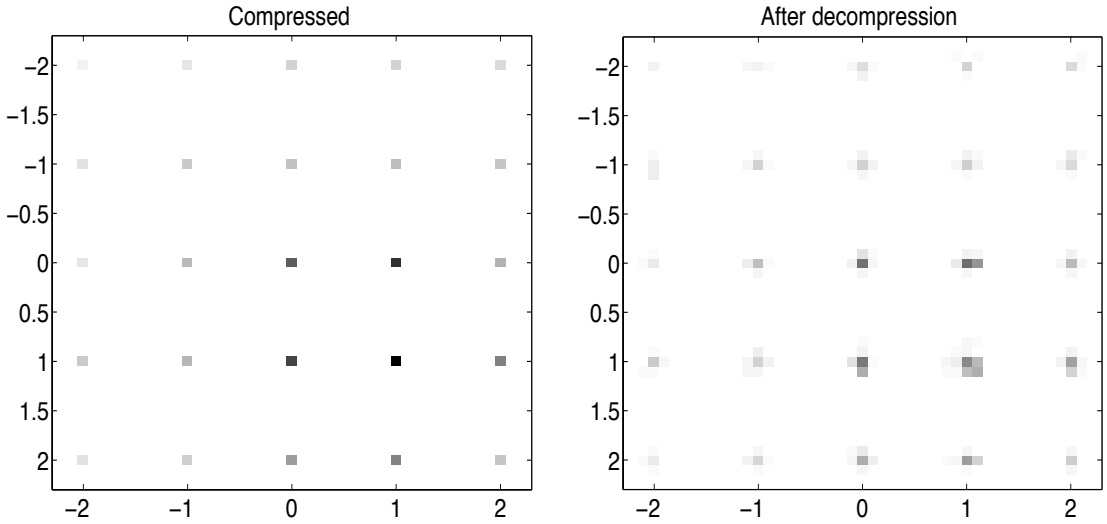

Figure 4.10: On the left is an empirical matrix of DCT coefficients after quantization. When decompressed to the spatial domain and rounded to pixel values, right, the DCT coefficients are randomly distributed around the quantization points.

If the image is re-compressed with a different quality level (i.e. different quantization step size), these blurred coefficients are rounded to the nearest new quantizer output. In some special cases, the first quantizer output value lies halfway between two output levels of the new quantizer. For example, if the first quantizer used a step size of 21, and the second quantizer uses 24, then $4 \times 21 = 84$ is straddled by $3 \times 24 = 72$ and $4 \times 24 = 96$. Since it is assumed that the distribution is white Gaussian, and therefore symmetric, it is expected that under normal quantization roughly half of the coefficients originally quantized to 84 become

72, and half 96. For pairs of coefficients, a quarter of pairs originally at (84,84) become (72,72), a quarter (72,96), a quarter (96,72) and a quarter (96,96) (see Figure 4.11). Fridirch et al propose changing the quantization of these values to add hidden data. If instead a value originally at 84 becomes 72 to represent a zero, and becomes 96 to represent one, the statistics are not expected to change.

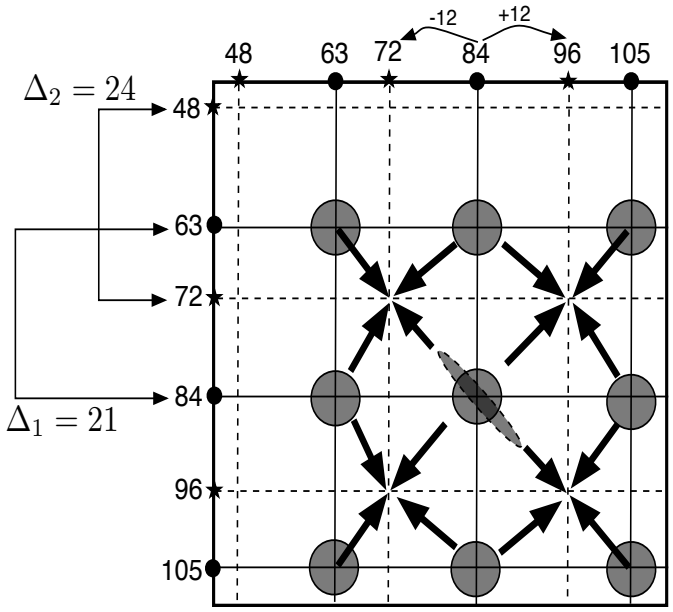

Figure 4.11: A simplified example of second compression on an empirical matrix. Solid lines are the first quantizer intervals, dotted lines the second. The arrows represent the result of the second quantization. The density blurring after decompression is represented by the circles centered at the quantization points. For the density at (84,84), if the density is symmetric, the values are evenly distributed to the surrounding pairs. If however there is an asymmetry, such as the dotted ellipse, the new density favors some pairs over others (e.g. (72,72), (96,96) over (72,96), (96,72). The effect is similar for other splits such as (63,84) to (72,72) and (72,96).

This statistical equivalency only fails if the density blurring is not, in fact, symmetric about the original quantization point. Though asymptotically it is expected to be, each realization is slightly asymmetric, as can be seen in Figure 4.10. We have found the asymmetry to be small, however the calculated divergence between a double compressed cover image and a PQ stego image is expected to be greater than zero. The net effect however is minimal and the divergence and detection results above are not surprising. Again we see a match between analysis, divergence measurements, and practical detection.

4.5 Outguess

Outguess is a data hiding tool designed to hide in the least significant bits of JPEG coefficients. Outguess 0.2 [68] is an improved version in which some of the coefficients are used for hiding, while the rest are used to restore the histogram to the cover state. Naturally this will prevent detection by histogram based techniques such as the chi-squared and approximate LLRT methods explored in Section 3.2. However it is shown by Fridrich et al [34] that although the histogram of DCT coefficients is compensated, the dependencies in the spatial domain will be changed. Specifically, the disruption at the boundaries of the 8 by 8 JPEG blocks is greater for Outguess hiding than for standard JPEG compression. We

note in Section 4.2.2 that the blockiness statistic used to measure this disruption is a function of the empirical matrix of the values straddling the JPEG blocks. A key component of detection done by Fridrich et al using the blockiness measure is a method of calibrating the statistic for each image, that is, a means a estimating the cover blockiness. They use a method designed for blockwise transform hiding schemes to do this. An estimate of the cover image is made by cropping the image under scrutiny by half the size of a JPEG block and re-compressing. From this estimated cover image, an estimate of the original blockiness is easy to compute.

We examine the detection of Outguess using the supervised learning detector exploiting spatial dependencies we use to detect spread spectrum hiding. That is, like Fridrich et al, we use spatial dependencies to detect Outguess. However we do not calibrate on a per image basis, but instead train on hundreds of examples. In Figure 4.12 we present results on various classifiers detecting Outguess in images. The feature vector is built similar to those used in Section 4.3.3 to detect spread spectrum hiding. To capture the block boundary dependency changes, vertical and horizontal dependencies are included in the feature. Detection here, as reflected in the ROC, is not particularly powerful. This demonstrates the benefit of using auto-calibration on an image-by-image basis. However such methods are not always available, and a possible solution for the steganographer is to simply use different block sizes.

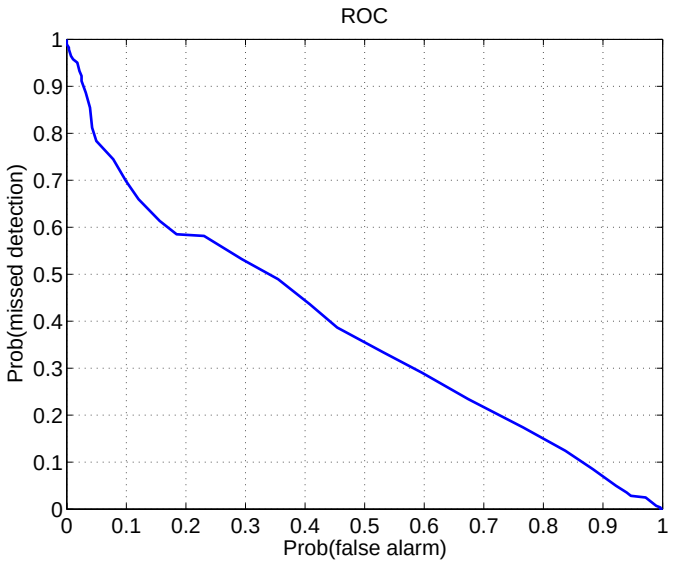

Figure 4.12: Detector performance of Outguess using classifier trained on dependency statistics.

4.6 Summary

Our Markov model for cover data permits explicit computation of a detection-theoretic divergence measure that characterizes the susceptibility of a steganographic scheme to detection by an optimal classifier. This measure has advantages over other steganographic security benchmarks. It provides a more accurate security measure than Cachin's ϵ-secure [10] metric, as dependencies between samples are accounted for. Additionally it is a more general metric than that given by Chandramouli et al [12], which is measured for a given detector. The divergence measure also provides a quick estimate of the performance benefits of using de-

pendency in steganalysis: the ratio of the divergence for the Markov model, to the divergence between marginal PMFs, represents the factor by which the use of dependency reduces the number of samples required for a given performance, relative to steganalysis based on one-dimensional histograms.

While the Markov model does not completely capture interpixel dependencies in images, we have shown it to be consistent with many image steganalysis schemes exploiting memory, which typically use a function of the statistics used to optimally discriminate between Markov source models. Furthermore, the detection-theoretic benchmarks computed using the Markov model are close to the performance attained by practical image steganalysis techniques. However, further research is needed into whether more complex statistical models can yield better image steganalysis techniques, and how to compute performance benchmarks for such techniques.

Improved models for images could include more degrees of dependency, as well as some model of non- or piecewise-stationarity. Recently we have seen work extend the i.i.d. approach to LSB analysis by employing models with memory: Draper et al [23] generalize our approach by considering adjacent pixel dependencides and Sidorov has performed hidden Markov model (HMM) analysis [78] on LSB hiding and used a MC derived statistic for LSB detection [79, in Russian]. Draper et al concluded that adding one level of dependency (i.e. considering two-

dimensional joint statistics) provided diminishing marginal returns. They observe that detectors employing a non-PMF statistic, such as RS analysis [33], perform very well, and consider PMFs an inferior tool for practical steganalysis. We stress that PMF analysis provides bounds on optimal detection, however these bounds are for the asymptotic case. Their observations are for finite a number of samples, an important consideration in comparing optimal detection to practical schemes. Sidorov found that considering one level of dependency (MC model) only showed good results for high-quality and synthetic images; for low-quality images the detector failed. In the related paper [78] the author considers extending the MC model to a Markov random field model, incorporating a higher degree of dependency, but does not implement such a detector. We note that the parameters of more complex models are more difficult to estimate for finite number of samples, a well-known problem [15]. Additionally variations from image to image may make it difficult to calibrate steganalysis techniques based on such models.

The Markov chain model can be related to Sallee's [75] model based approach. Rather than assuming the steganalyst is limited to a parameterized distribution model, we assume the actual joint distribution is used. Just as with Sallee's steganography, the security is only as good as the model. Thus, much work remains to be done on the fundamental problem of understanding how the complexity of the model for the cover data impacts the accuracy of estimating the model

parameters, and the computational complexity and performance of steganalysis based on the model.

Chapter 5

Evading Optimal Statistical Steganalysis

We now turn our attention to applying our detection-theoretic framework to *prevent* detection. In particular we seek to find methods that maintain an adequate rate, and can withstand at least nominal interference between sender and receiver, for example from an active warden, transmission noise, or compression. We note many previous methods designed to reduce detectability (outlined in Chapter 2) have focused on a particular steganalysis statistic. We know from our study of steganalysis that for perfectly secure embedding, the distribution of the stego data must exactly match the cover, which we have seen does not occur by default. Other methods such as stochastic modulation [31], JPEG perturbed quantization [38], and model based embedding [75, 76] accept a change of stego distribution from the original cover, but attempt to match a different distribution, which is close to a plausible cover distribution. It is difficult to define what is

"plausible enough", and in some cases (e.g. [7]) a steganalyst can exploit the divergence from the original. Additionally, these approaches are very fragile to any interference between sender and receiver.

Provos' Outguess [68] is an early attempt at restoring the stego distribution to the cover empirical distribution. This method was followed up later by Eggers et al [26], with a more mathematical formulation denoted histogram-preserving data mapping (HPDM), and Franz, with work in matching the message data to the cover distribution [29]. All of these schemes are designed for compensating discrete-valued hiding medium. Theoretical proposals for continuous-valued hiding compensation have also been presented. In [43] Guillon et al suggest transforming the cover samples before hiding, such that the cover samples are uniformly distributed. QIM hiding in uniformly distributed data does not change the distribution. It is likely however that robustness would suffer with such a scheme, and it may be difficult to implement in practice. Wang and Moulin [98] proposed stochastic QIM, which achieves zero K-L divergence. Because of the random nature of the hiding, it is difficult to judge the robustness of stochastic QIM.

5.1 Statistical Restoration Scheme

As with our study of steganalysis, we begin with an i.i.d. assumption and later extend this to a Markov chain model. Under the independence assumption, our goal for perfect security is to match the marginal distribution. In practice, the underlying distribution of an image is not known. Our approach then is to match the empirical distribution of the image. Since the empirical distribution is derived from a cover image, there is no doubt it is a plausible cover distribution. The plan is to save coefficients otherwise used for embedding to instead compensate the distribution, similar to Provos' [68] method to restore the histogram of JPEG coefficients. The goal here however is to recover a continuous distribution, the pdf, rather than the PMF. In this way, hiding media such as DCT coefficients which have not been quantized can be statistically compensated. Naturally, if these coefficients are later quantized, they will still match the expected distribution of the quantized cover.

Practically speaking, the steganalyst does not have access to continuous pdfs, but instead calculates a histogram approximation. Our data hiding is secure if we match the stego histogram to the cover histogram using a bin size, denoted w, the same size as, or smaller than, that employed by the steganalyst. We stress that all values are present and there are no "gaps" in the distribution of

values; however, within each bin the data is matched to the bin center. A key assumption is that for small enough w, the distribution is uniformly distributed over the bin, a common assumption in source coding [40]. Under this assumption, we can generate uniformly distributed pseudorandom (PR) data to cover each bin, and still match the original statistics. Let $f_X(x)$ be the cover probability density function (pdf) and $f_S(s)$ the stego pdf. For I bins centered at $t[i]$, $i \in [1, I]$ with constant width w, the expected histogram for data generated from $f_X(x)$ is:

$$P_X^E[i] = \int_{t[i]-w/2}^{t[i]+w/2} f_X(x)dx \tag{5.1}$$

with a similar derivation of $P_S^E[i]$ from $f_S(s)$. The superscript E denotes that this is the expected histogram, to discriminate it from histograms calculated from random realizations. We generally refer to these expected quantized pdfs as PMFs.

The hiding scheme we analyze here is outlined in greater detail in [81]. The key points we need for our analysis are as follows:

- A proportion of coefficients are used for hiding, the rest are used for compensating the distribution after hiding.

- The compensating coefficients are uniformly (PR) distributed within the bins, so there are no gaps in the pdf. In our QIM implementation this is accomplished with dithered quantization.

- A minimum mean squared error (MMSE) criteria is used to minimize distortion while hiding

Let $\lambda \in [0, 1)$ be the ratio of symbols used for hiding, so $1 - \lambda$ is the ratio remaining to match the cover histogram. If $P_X[i]$ is the cover histogram, $P_S[i]$ the standard (uncompensated) stego histogram, $P_C[i]$ the histogram of compensating coefficients, and $P_Z[i]$ the histogram of the final output, our goal can be summarized as:

$$\begin{aligned} P_Z[i] &= P_X[i] \; \forall \, i \\ P_Z[i] &\triangleq \lambda P_S[i] + (1 - \lambda) P_C[i] \end{aligned} \tag{5.2}$$

We first examine the effect of allowing a small, low probability region to remain uncompensated after hiding. By ignoring these regions, the rate can be increased significantly. However since the distribution is not perfectly matched, there is a measurable K-L divergence between the cover and stego pdfs. We study this tradeoff between divergence and rate, and find, not surprisingly, that restoration effects a more efficient tradeoff than simply embedding in fewer coefficients. We then focus on eliminating all divergence, and derive an expression for the expected rate with a given cover distribution. Finally we expand this analysis to statistical restoration of the joint histogram, or empirical matrix.

5.2 Rate Versus Security

Using a set of embedding coefficients for compensating instead of hiding reduces the size of the message that can be hidden; this is the cost of increasing security. We can characterize this cost by studying the amount of data that can be hidden in an idealized data source with a given distribution.

The amount of data that can be hidden is proportional to the number of symbols that can be hidden in. So to maximize the amount of data we send, we seek to maximize λ for a given cover histogram subject to the constraint in (5.2), and the constraints imposed on the distribution of compensating coefficients, namely $\sum P_C[i] = 1$ and $P_C[i] \geq 0 \; \forall \, i$. The first constraint is true for all λ. For the second constraint, by substituting an expression for P_C satisfying (5.2), we find:

$$\begin{aligned} P_C[i] &\geq 0 \; \forall \, i \\ \frac{P_X[i] - \lambda P_S[i]}{(1-\lambda)} &\geq 0 \\ -\lambda P_S[i] &\geq -P_X[i] \\ \lambda &\leq \frac{P_X[i]}{P_S[i]} \; \forall \, i \end{aligned}$$

which gives us an upper limit on the percentage of symbols we can use for hiding.

Ideally the sender would use a different λ for each bin. However hiding shifts values from one bin to another, and the valid λ for one bin may be larger than an adjacent bin and the constraint is violated. Because of this, a worst-case λ is

chosen:

$$\lambda^* \triangleq \min_i \frac{P_X[i]}{P_S[i]}$$

However if we apply this constraint on λ to typical PMFs we run into erratic behavior in the low-probability tails. The ratio $\frac{P_X[i]}{P_S[i]}$ can vary widely here, from infinitesimally small to quite large. This is because the magnitude of changes to the ratio depends on the value of $P_X[i]$, for example: $\frac{0.50}{0.49}$ is very close to $\frac{0.50}{0.51}$ whereas $\frac{0.02}{0.01}$ is quite different from $\frac{0.02}{0.03}$. We study this phenomenon in greater detail in Section 5.3. We note however that since the probabilities of the corresponding events are low, then the effect of PMF differences in these regions on the net divergence is small. So to avoid this problem we can relax the exact equality constraint and not require compensation in a small, low probability region of the PMF. In the next section, we focus on achieving perfect matching for the entire PMF. With the relaxed constraint:

$$\lambda^* = \min_{i \in \mathfrak{C}} \frac{P_X(i)}{P_S(i)}$$

where $\mathfrak{C}$ is the compensated region. In addition to the divergence introduced to the ignored ($\mathfrak{C}^c$) region, since (5.2) is no longer true for all i, P_C must be normalized to satisfy the unity sum constraint, adding a small change across the PMF. Though the net effect is to introduce a small amount of divergence, λ and the corresponding hiding rate can only increase. A tradeoff can be made between

desired security from detection and the hiding rate. We note that another method to increase security at the cost of embedding payload is to simply embed in fewer coefficients. However we expect that explicitly fixing the histogram with the remaining coefficients has a greater effect at reducing divergence. We verify this assumption for a specific scheme outlined below.

To gauge the effect of hiding in practice, we examine the compensation scheme adapted to QIM hiding with dithering (see Section 3.3). We examine the hiding effect on covers with Gaussian and Laplacian distributions with different variances. Additionally we change hiding parameters such as the step size of the hiding quantizer, Δ. As expected from our findings on QIM steganalysis, we found the rate and divergence to be related to the ratio σ/Δ within a given PMF family. For our tests comparing rate and divergence, we would then expect different tradeoffs for different cover distributions and different σ/Δ.

To find the rate-divergence tradeoff, we find the rate corresponding to several different sizes of ignored (uncompensated) regions. A larger ignored region has a greater divergence from the original, but eliminates small $\frac{P_X[i]}{P_S[i]}$, which increases λ, the hiding rate. This tradeoff for restoration is compared to the tradeoff for simply embedding less. In Figure 5.1 we see a large decrease in divergence can be made with a small drop in rate, which is not possible by merely embedding less. This is true for both Laplacian and Gaussian PMFs over a range of σ/Δ.

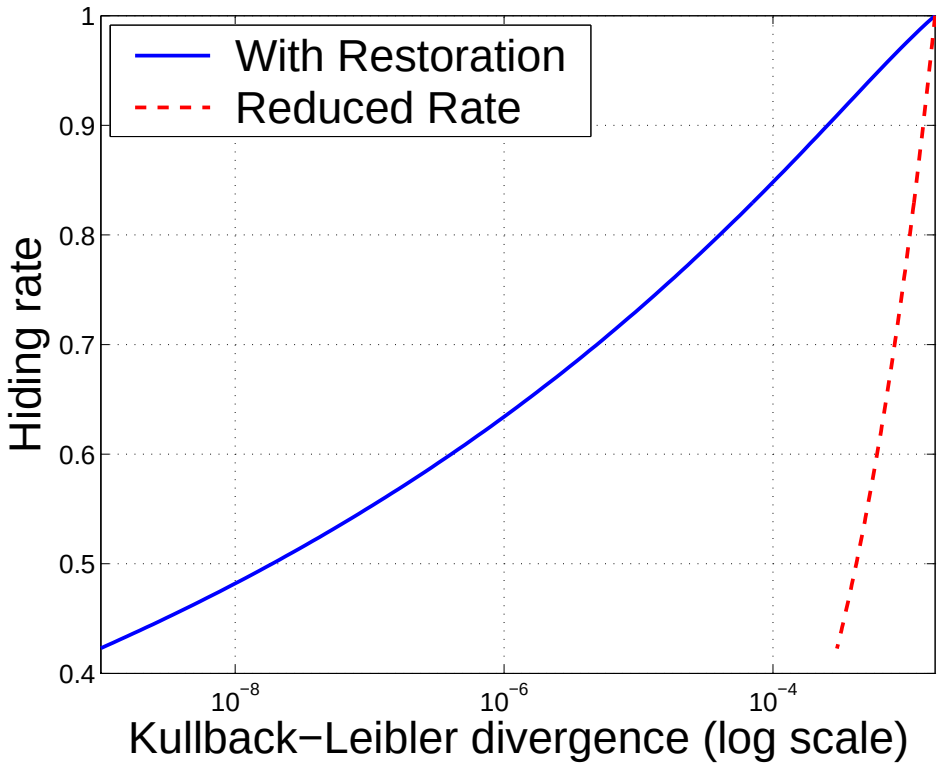

Figure 5.1: Rate, security tradeoff for Gaussian cover with σ/Δ of 1. As expected, compensating is a more efficient means of increasing security while reducing rate.

5.2.1 Low Divergence Results

To examine the efficacy of statistical restoration, we compare the divergence between cover and stego for standard hiding and for hiding with compensation at the same rate. For Gaussian PMFs at a rate of $\lambda = 0.35$ the divergence for standard hiding is 5.9×10^{-3}, and 1.3×10^{-3} for compensated. The standard divergence is nearly five times greater. For a set of real image statistics, hiding

at the same rate, the average divergence for standard hiding is 6.5×10^{-3}, and 2.1×10^{-3} for compensated. Although detection is still possible, since divergence is greater than zero, restoration greatly increases the error probabilities of an ideal detector. For example, a steganalyst would require more than three times as many samples to achieve the same detection rates with standard hiding in images as with hiding with restoration.

Due to this decrease in divergence, we would expect a steganalysis scheme to perform better detecting standard hiding compared to compensated hiding. We compared the detection rates for using the practical steganalysis scheme used in Chapter 3 to detect standard dithered QIM hiding, and an image adaptive dithered QIM scheme [81]. We trained and tested two machines on the same sets of images and at the same rate; one with restoration, one without. The test used a histogram of DCT coefficients as a statistic. The results are summarized in Table 5.1.

	Std. Dithered QIM		Adaptive Dithered QIM	
	Uncomp.	Comp.	Uncomp.	Comp
P(Miss)	0.075	0.525	0.701	0.796
P(False Alarm)	0.177	0.000	0.000	0.074
P(M. +F.A.)	0.252	**0.525**	0.701	**0.870**

Table 5.1: It can be seen that statistical restoration causes a greater number of errors for the steganalyst. In particular for standard hiding, the sum of errors for the compensated case is more than twice that the uncompensated.

We have found that detection rates are decreased, but are still better than random guessing, which is expected due to the allowed difference between cover and stego distributions. We now examine zero divergence hiding and rate.

5.3 Hiding Rate for Zero K-L Divergence

We here analyze the proposed solution for perfect restoration and characterize the maximum message size that can be hidden. In particular, by finding the distribution of the rate allowing perfect restoration, we can find a rate guaranteeing zero divergence within a given probability threshold. Though the approach we use can be applied to statistical restoration of any hiding scheme, we show the specific example of our QIM scheme.

5.3.1 Rate Distribution Derivation

We now examine an approach where no region is left uncompensated, and so the histogram can be perfectly matched. To do this, we need to find the distribution of the minimum of the histogram ratio, λ^* for a given cover pdf, $f_X(x)$. We note that histograms calculated from real data $\{X_n\}_1^N$ vary for each realization. In other words, the number of symbols in the bins, $NP_X[i]$, are random variables. Be analyzing the distribution of these random variables, we

can find the distribution of the ratio $\frac{P_X[i]}{P_S[i]}$ for each bin, and from this the statistics of λ^*. We point out that the density of a ratio is not equal to the ratio of the density, that is, in general

$$f\left(\frac{P_X[i]}{P_S[i]}\right) \neq \frac{f(P_X[i])}{f(P_S[i])}$$

Let $V_X[i] = NP_X[i]$ be the number of symbols from $f_X(x)$ falling into bin i, then $V_X[i]$ has binomial density function $P_{V_X[i]} = B\{N, P_X^E[i]\}$ [77]. Similarly if $V_S[i]$ is the number of symbols per bin for data from $f_S(s)$, it is distributed as $B\{N, P_S^E[i]\}$. See Figure 5.2 for a schematic of finding the distribution of the bins of a histogram.

We now define $\Gamma[i] \triangleq \frac{V_X[i]}{V_S[i]} = \frac{P_X[i]}{P_S[i]}$. The cumulative distribution of $\Gamma[i]$, $F_{\Gamma[i]}(\gamma) = P(\Gamma[i] \leq \gamma)$, is given by

$$F_{\Gamma[i]}(\gamma) = \sum_{k=0}^{N} \sum_{l=0}^{\lfloor \gamma k \rfloor} P_{V_S[i]}(k) P_{V_X[i]}(l)$$

and the density is

$$f_{\Gamma[i]}(\gamma) = \frac{dF_{\Gamma[i]}(\gamma)}{d\gamma}.$$

Ultimately, we wish to find the distribution of the minimum Γ over all bins, giving us a statistical description of λ^*, our zero-divergence hiding rate. The cumulative distribution of λ^* is the distribution of $\min_i \Gamma[i]$ given by

$$F_{\lambda^*}(\gamma) = 1 - \left\{ \prod_i \left[1 - F_{\Gamma[i]}(\gamma)\right] \right\}$$

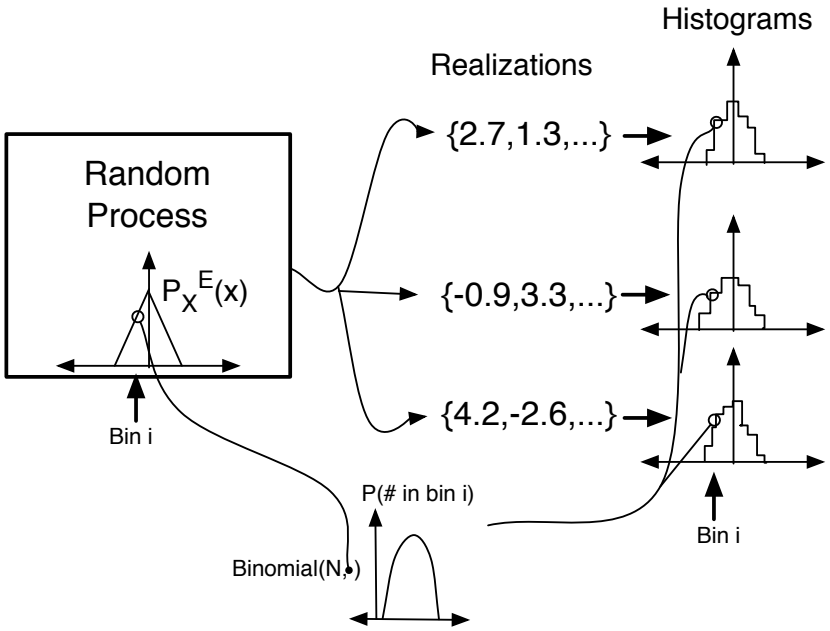

Figure 5.2: Each realization of a random process has a slightly different histogram. The distribution of the number of elements in each bin is binomially distributing according to the expected value of the bin center (i.e. the integral of the pdf over the bin).

[65, Sec 7.1] and the density can be found by differentiating. To summarize, given the pdfs of cover and stego, $f_X(x)$ and $f_S(s)$, we can find the distribution of λ^*: the proportion of symbols we can use to hide in and still achieve zero divergence. Using this, the sender and receiver can choose ahead of time to use a fixed λ that guarantees zero-divergence (i.e. $\lambda \leq \lambda^*$) within a desired probability. In Section 5.3.3 we illustrate this analysis with an example, but first we examine the factors affecting the rate.

5.3.2 General Factors Affecting the Hiding Rate

By examining the derivation of the distribution of λ^*, we can predict the effect of various parameters on the hiding rate. The key factors effecting the payload are:

1. **Cover and stego pdfs,** f_X, f_S: Obviously the "closer" the two pdfs are to one another, the less compensation is required, and the higher the rate. The difference between the pdfs generally depends on the hiding scheme.

2. **Number of samples,** N: The greater the number of samples, the more accurate our estimates of the samples per bin. Therefore it is easier to guarantee a λ to be safe with given probability, and so the hiding rate is higher. The number of samples is mostly a function of the size of the image.

3. **Bin width,** w**, used for compensation**: Bin width is important to guaranteeing security, but the effect of bin width is not immediately clear, and is a research topic in its own right [77, 95]. We briefly explore the net effect of w on λ^* below.

To judge the effect of w, we examine the expected value of Γ, the ratio of P_X to P_S. To avoid problems with dividing by zero we instead use $\Gamma' \triangleq \min(\Gamma, 1)$. Since Γ greater than one implies embedding at greater than 100%, these cases can be ignored. From (5.1) we note that a smaller bin width leads to smaller expected

bin values, $P_X^E[i]$ and $P_S^E[i]$. By evaluating

$$E\{\Gamma'\} = \sum_{k=0}^{N} \sum_{l=0}^{n} \min\left(\frac{l}{k}, 1\right) P_{V_S[i]}(k) P_{V_X[i]}(l)$$

for several distributions $P_{V_S[i]}, P_{V_X[i]}$ (recall these are binomially distributed based on P_S^E, P_X^E), we find that for a set ratio of expected bin values, $P_X^E[i]/P_S^E[i]$, the expected value of Γ' increases as P_S^E and P_X^E decrease. In other words, the expected value of Γ' increases as the bin width decreases. This is particularly true for small P_X^E (and therefore small Γ'). For larger P_X^E (greater than 0.1) the expected value of the ratio is basically equal to the ratio of the expected value for any bin size. Additionally, the variance of Γ' (evaluated in the same manner as the mean) decreases as P_S^E and P_X^E decrease. Both of these trends would indicate that decreasing w increases the hiding rate. However this only accounts for a reduction of bin width for a fixed bin center. As w decreases, new bins must be created. λ^*, is evaluated as the minimum over all the bins. Intuitively, as the number of bins increases there is an increased probability of a smaller minimum.

So the net effect, an increase or decrease in $E\{\lambda^*\}$, depends on the distributions. We then must evaluate the net effect of bin width on λ^* for given cover and stego pdfs. Fortunately for the steganographer, the steganalyst can not choose an arbitrarily small bin size in order to detect, as the mean integrated square error (MISE) of the detector's estimate of the pdf is not simply inversely related to bin width [77]. In other words, the detector also faces a challenge in choosing an ap-

propriate bin size. Empirically, both the steganographer and steganalyst should try to use as large a bin width as possible, while still maintaining the uniform approximation over each bin.

5.3.3 Maximum Rate of Perfect Restoration QIM

We now apply the findings of the previous section to the case of quantization index modulation (QIM) data hiding, specifically dithered uniform scalar QIM [14]. For a given cover pdf $f_X(x)$ we know the expected pdf for data after undergoing dithered scalar QIM with constant step-size Δ from our steganalysis of QIM (Section 3.3):

$$f_S(s) = \frac{1}{\Delta}\int_{s-\Delta/2}^{s+\Delta/2} f_X(x)dx \equiv \left[f_X(x) * \frac{1}{\Delta}\sqcap(x/\Delta)\right](s)$$

where $\sqcap(t)$ is the rectangle function, defined in terms of the unit step function $u(t)$ as $\sqcap(t) = u(t+1/2) - u(t-1/2)$. As we found in Section 3.3, without compensation, this change can be detected when hiding in real images. We now show how to find λ^* allowing for perfect histogram restoration of QIM hiding.

Earlier we listed general factors affecting the amount of data that can be embedded while allowing histogram compensation. In the context of QIM hiding we can more explicitly characterize these factors. Since we can calculate f_S from f_X, of this pair we need only examine the cover pdf. Since f_S is a smoothed

version of f_X (after convolution), the ideal f_X for secure QIM hiding is uniform [43], which can not be smoothed any further. Unfortunately it is difficult to find a uniformly distributed cover medium. Typical hiding medium, particularly transform domain coefficients, are sharply peaked. As noted earlier, σ/Δ is an important parameter for QIM. For large σ/Δ, the cover pdf is flat relative to the quantization interval, and less change is caused to the original histogram by hiding. Therefore less samples are needed for compensation, and more are available for hiding, and the expected λ^* is large. Conversely a small σ/Δ has a low expected λ^*.

Of all the factors, only w and Δ are in the hands of the steganographer. Decreasing Δ increases σ/Δ, and therefore the safe hiding rate. However, decreasing Δ also increases the chance of decoding error due to any noise or attacks [14]. Thus if a given robustness is required, Δ can also be thought of as fixed, leaving only the bin width. For QIM hiding in Gaussians and Laplacians, we found that decreasing the bin size w led to a decrease in λ^*, suggesting that the steganographer should choose a large w. However, as mentioned before, w should be chosen carefully to avoid detection by a steganalyst using a smaller w.

We presently examine an idealized QIM scheme, followed by an extension to our practical QIM scheme which prevents decoder errors. As an illustrative ex-

ample, we provide results derived for hiding in a Gaussian, i.e. $f_X(x) = \mathcal{N}(0, \sigma_X^2)$, but note the approach can be used for any $f_X(x)$.

In Fig 5.3 is the density of $\Gamma[i]$, $f_{\Gamma[i]}(\gamma)$ for all i and a range of γ, for QIM hiding in a zero mean unit variance Gaussian. From this density we can see the

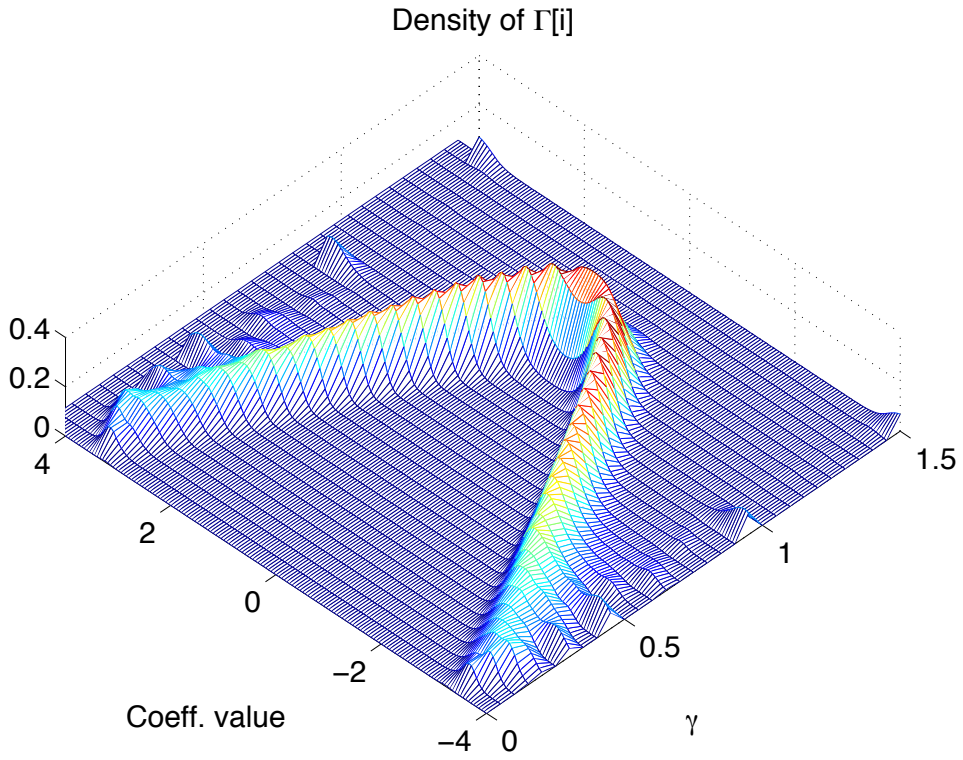

Figure 5.3: The pdf of Γ, the ratio limiting our hiding rate, for each bin i. The expected Γ drops as one moves away from the center. Additionally, at the extremes, e.g. ± 4, the distribution is not concentrated. In this example, $N = 50000$, $\sigma/\Delta = 0.5$, and $w = 0.05$.

relationship between Γ and bin center. For bins located near zero, $\Gamma[i]$ has a probability concentrated above 1 (though obviously we can not embed in more

than 100% of the coefficients). For bins a bit further from the center, the expected value for Γ drops. Since the effect of dithered QIM is to smooth the cover pdf this result is not surprising. The smoothing moves probability from the high probability center out towards the tails, see for example Fig. 5.4. Though this

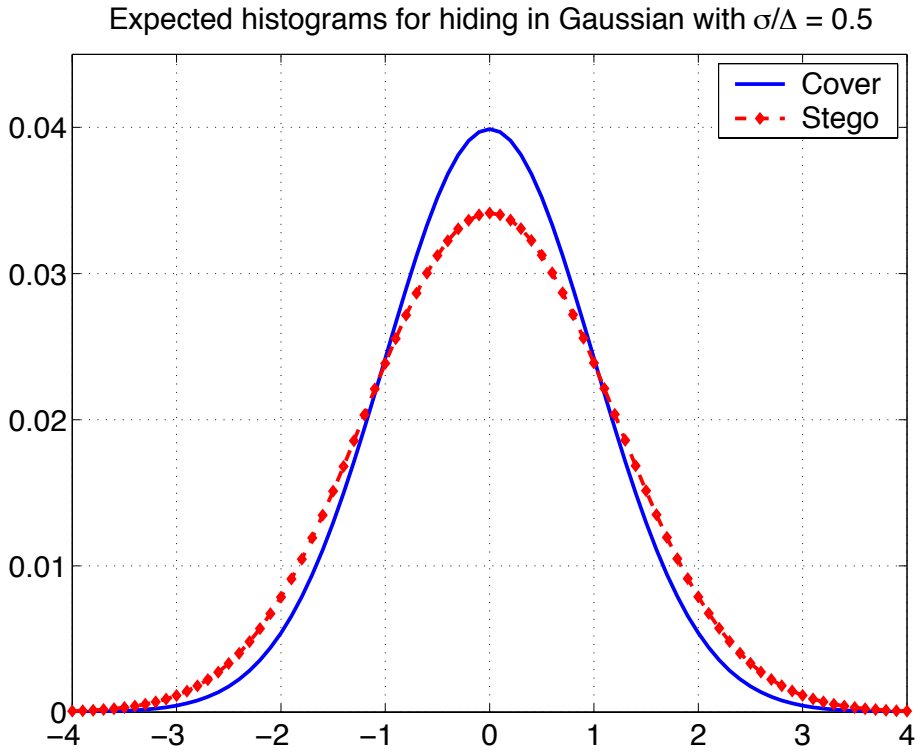

Figure 5.4: The expected histogram of the stego coefficients is a smoothed version of the original. Therefore the ratio $\frac{P_X^E[i]}{P_S^E[i]}$ is greater than one in the center, but drops to less than one for higher magnitude values.

result is found for hiding in a Gaussian, we expect this trend from any peaked unimodal distribution, such as the generalized Laplacian and generalized Cauchy distributions often used to model transform coefficients [83, 80, 75]. Near the ends,

e.g. ± 4, Γ is distributed widely over all possible values. So while it is possible to have a very high γ here, it is also possible to be very low; i.e. the variance is very high. We empirically observed these problems in the low probability tails in Section 5.2. The first solution proposed is to hide in all coefficients but only compensate bins away from these tails. We now attempt to only *hide* in the high probability region; after hiding, only this region needs to be compensated. This introduces a practical problem, the decoder is not always able to distinguish between embedded coefficients and non-embedded. We address this issue below, but first we examine the ideal case. So our motivation is, even though we must reduce the number of embeddable coefficients by not embedding in high valued, low probability coefficients, our net rate may be higher due to a higher λ^*, where λ^* is redefined as

$$\lambda^* \triangleq \min_{i \in \mathcal{H}} \frac{P_X[i]}{P_S[i]}$$

where $\mathcal{H}$ is the hiding region, defined as $\mathcal{H} \triangleq [-T, T]$ where T is a threshold. The net hiding rate, no longer simply equivalent to λ^*, is now

$$R = \lambda^* G(\mathcal{H})$$

where

$$G(\mathcal{H}) \triangleq \sum_{i \in \mathcal{H}} P_X[i].$$

As the threshold T, increases, more coefficients are available for embedding, as seen in Fig. 5.5. However as the threshold is increased the expected λ^* decreases, resulting in a lower rate. Practically the encoder and decoder can agree on a λ which leads to perfect restoration within a pre-determined probability, 90% for example. From the distribution of λ^*, the λ guaranteeing perfect restoration with a given probability can be found for each threshold. These 90%-safe λs decrease as the threshold is increased, as seen in Fig. 5.6, along with an example of deriving the 90%-safe λ for the threshold of 1.3. The net effect of an increasing $G(T)$ and decreasing safe λ is a concave function from which the maximum rate can be found.

In Fig. 5.7 we show the relationship between the chosen threshold and the rate allowing perfect histogram matching in 90% of cases. In this case, the maximum rate is 0.65 bits per coefficient. So, using a threshold of 1.3 and a λ of 0.81 (from Figure 5.6), the hider can successfully send at a rate of 0.65, and the histogram is perfectly restored nine times in ten.

5.3.4 Rate of QIM With Practical Threshold

There is however a practical problem when implementing the scheme addressed in the above derivation. At the decoder, there is ambiguity with values near the threshold. In the region near the threshold, the decoder does not know if the

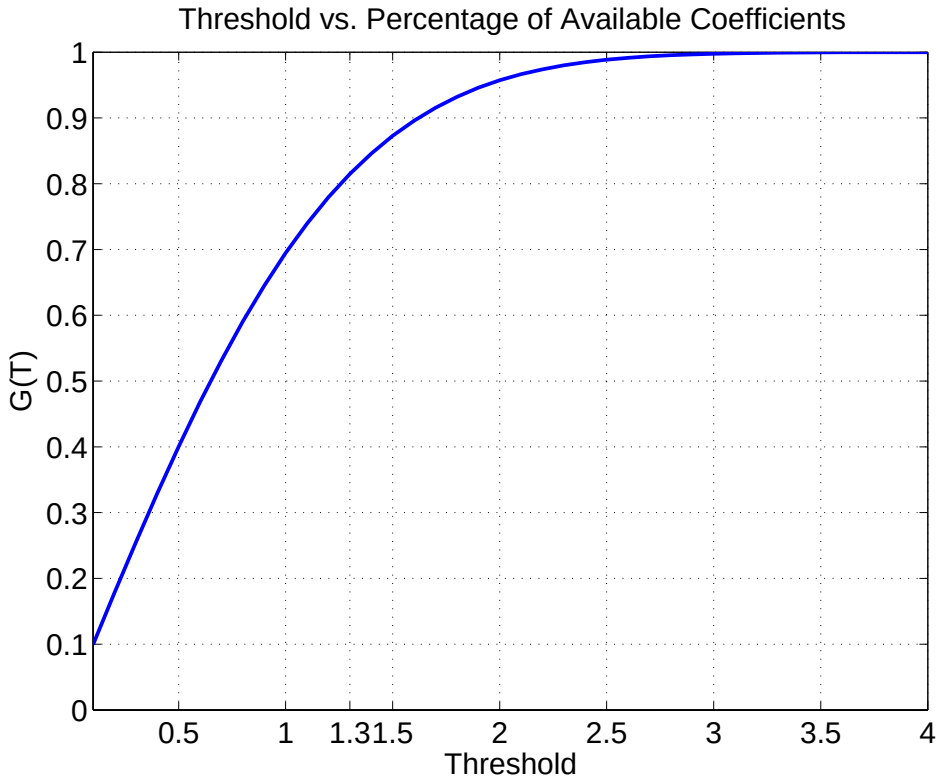

Figure 5.5: A larger threshold allows a greater number of coefficients to be embedded. This partially offsets the decrease in expected λ^* with increased threshold.

received value is a coefficient that originally was below the threshold and is now shifted above the threshold after hiding and dithering, or is simply a coefficient that originally was above the threshold and contains no data. Therefore we create a buffer zone near the threshold: if, after hiding, a coefficient would be shifted above the threshold, it is instead skipped over. To prevent creating an abrupt transition in the histogram at the buffer zone, we dither the threshold with the

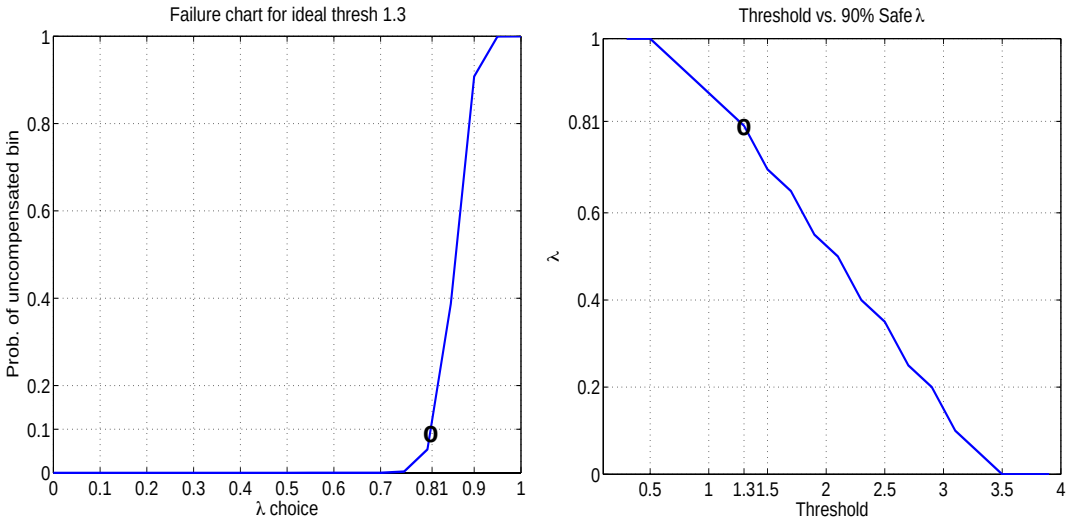

Figure 5.6: On the left is an example of finding the 90%-safe λ for a threshold of 1.3. On the right is safe λ for all thresholds, with 1.3 highlighted.

dither sequence. Since the decoder knows the dither sequence, this should not introduce ambiguity. This solution clearly results in a different stego pdf, $f_S(s)$, particularly near the threshold. Now we have

$$f_S(s) = \begin{cases} \frac{1}{\Delta}\int_{s-\Delta/2}^{s+\Delta/2} f_X(x)dx & s \in |s| < T - \Delta/4 \\ g(s)f_X(s) + \frac{1}{\Delta}\int_{s-\Delta/2}^{s} f_X(x)dx & s \in [T-\Delta/4, T+\Delta/4) \\ f_X(s) & \text{else} \end{cases}$$

where $g(s)$ is a scaling function:

$$g(s) = \begin{cases} |s| - T + \frac{\Delta}{4} & |s| \in [T-\Delta/4, T+\Delta/4) \\ 0 & \text{else} \end{cases}$$

So in the center region before the threshold, the stego pdf is the same as the previous case. In the region near the threshold, there is a blending of $f_X(x)$

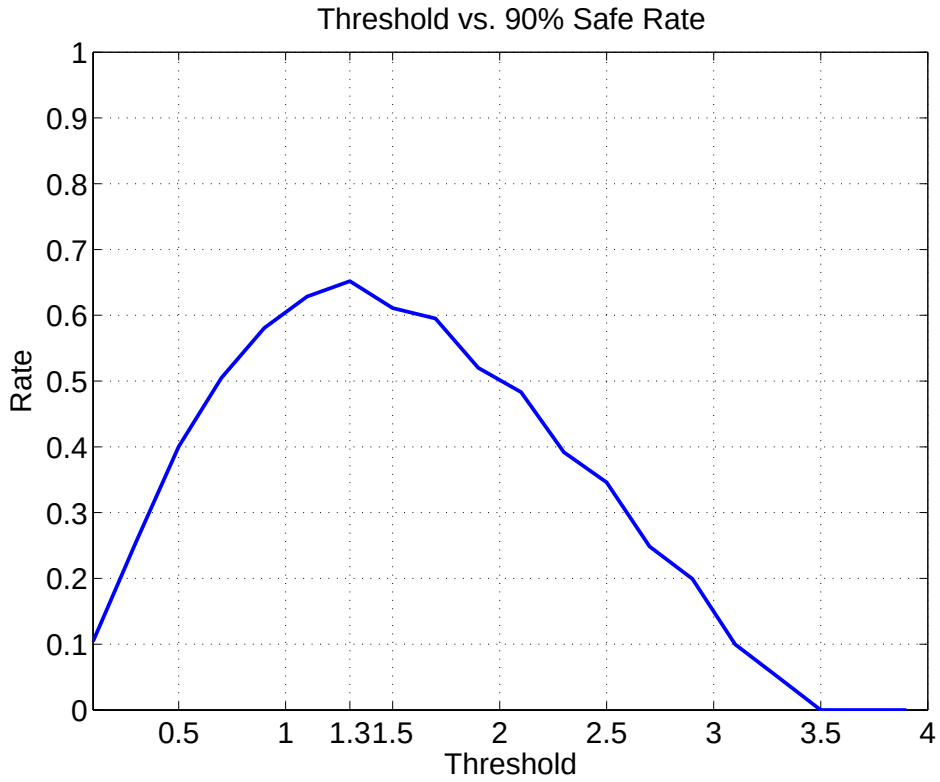

Figure 5.7: Finding the best rate. By varying the threshold, we can find the best tradeoff between λ and the number of coefficients we can hide in.

and a weakened (integrated over a smaller region) version of the standard $f_S(s)$. Finally beyond the threshold region, the original coefficients pass unchanged and the statistics are unaffected. In Fig. 5.8, we have f_X, f_S for the ideal case, and f_S for our practical threshold scheme, shown for two thresholds. Clearly $P_S^E(s)$ is higher in the threshold region than for the ideal case, so our rate is less than the ideal case.

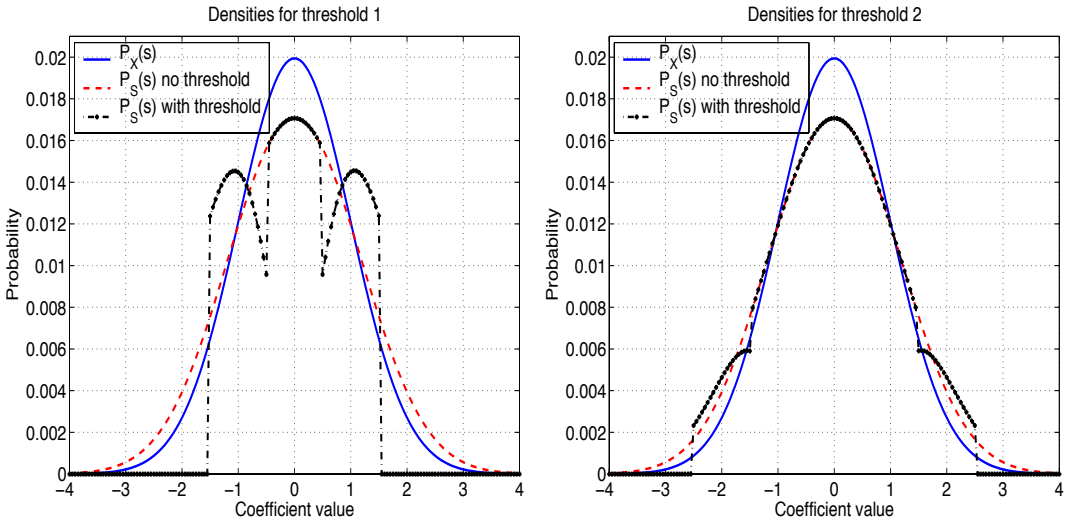

Figure 5.8: A comparison of the expected histograms for a threshold of one (left) and two (right). Though the higher threshold densitie appears to be closer to the ideal case, the minimum ratio P_X/P_S is lower in this case.

As with the ideal threshold case, we can calculate a λ guaranteeing perfect restoration a given percentage of the time. In Fig. 5.9, we show the pdf of Γ. The thresholding effect can be clearly seen: though the expected Γ is increased near the threshold, it drops quickly after this.

Finally Table 5.2 shows the 90%-safe rate for various thresholds. Here we would choose a threshold of 1, to achieve a rate of 0.3, about half the rate of the ideal case.

We have compared the derived estimates to Monte Carlo simulations of hiding and found the results to be as expected for different parameters $(n, w, \sigma/\Delta)$. We

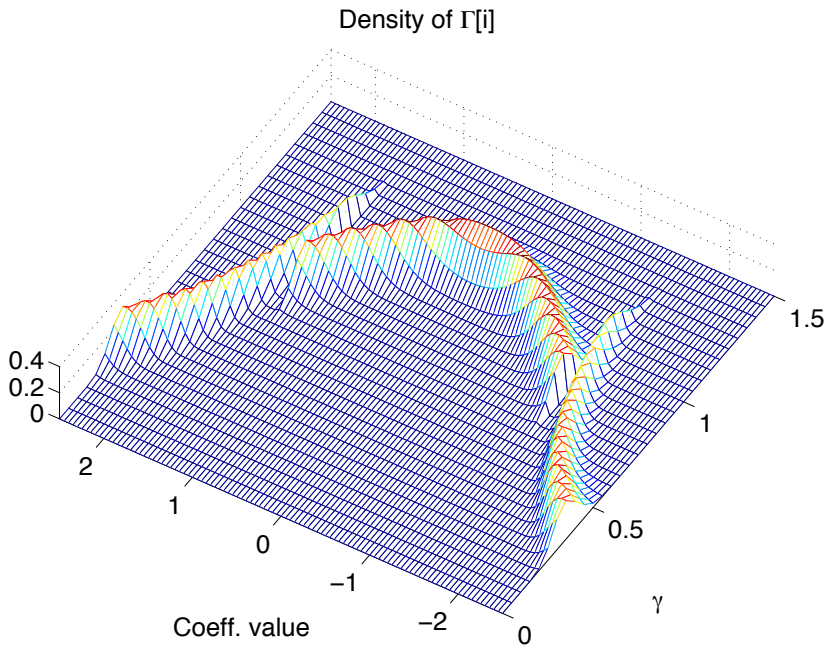

Figure 5.9: The practical case: Γ density over all bins within the threshold region, for a threshold of two. Though for bins immediately before the threshold, Γ is high, the expected Γ drops quickly after this. As before, $N = 50000$, $\sigma/\Delta = 0.5$, and $w = 0.05$.

therefore have an analytical means of prescribing a choice of λ and T for maximum hiding rate guaranteeing perfect restoration within a given probability.

5.3.5 Zero Divergence Results

We trained and tested an SVM classifier using a feature vector based on the first order statistics of DCT coefficients. To compare the perfect restoration scheme with standard hiding, we use both hiding methods to embed random data into hundreds of images. The same rate and the same images are used with

Threshold vs. Rate			
Threshold	1	2	3
$G(T)$	0.66	0.94	0.99
90%-safe λ	0.45	0.25	NA
Safe rate	0.30	0.24	0

Table 5.2: An example of the derivation of maximum 90%-safe rate for practical integer thresholds. Here the best threshold is $T = 1$ with $\lambda = 0.45$ There is no 90%-safe λ for $T = 3$, so the rate is effectively zero.

both schemes. As expected, perfectly restoring the histogram foils a detection method that can detect standard hiding at the same rate, see Figure 5.10. We note the effective hiding rate is lower than that used in our experiments in Section 5.2.1.

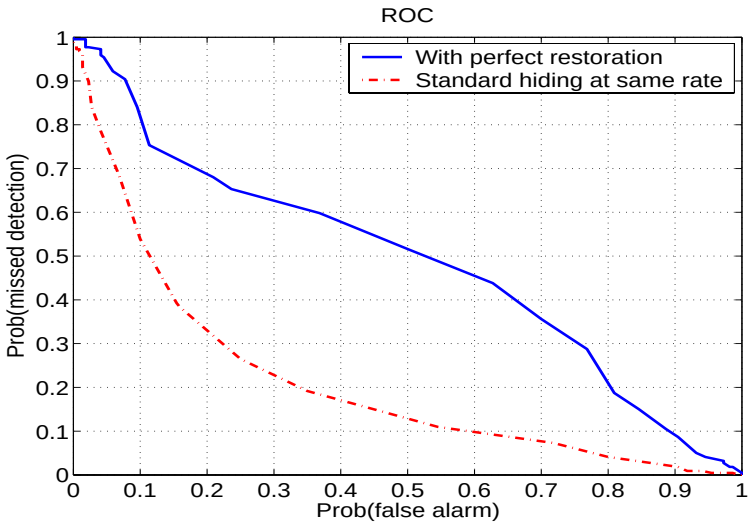

Figure 5.10: A comparison of practical detection in real images. As expected, after perfect restoration, detection is random, though non-restored hiding at the same rate is detectable.

5.4 Hiding Rate for Zero Matrix Divergence

We now turn our attention to avoiding steganalysis that uses memory. As in Chapter 4 we use a Markov chain model to include dependencies. To extend our statistical compensation scheme discussed above, we now seek to compensate the empirical matrix or joint histogram. In other words, we now consider changes to pairs of values, rather than individual values.

5.4.1 Rate Distribution Derivation

In practice, it is difficult to implement a scheme to compensate pairs throughout the entire image while still minimizing the mean square error introduced. On the other hand, extending our analysis of histogram restoration is straightforward. By doing so, we can estimate the best achievable hiding rate when considering joint statistics in an idealized setting. Let $f_{X_n,X_{n-1}}(x_n, x_{n-1})$ be the joint pdf of the cover, and $f_{S_n,S_{n-1}}(s_n, s_{n-1})$ the joint pdf of stego. As before, we assume the steganalyst is using a histogram approximation. The expected joint histogram is found similar to the one-dimension case (Eqn 5.1):

$$P^{(E)}_{X_n,X_{n-1}}[i,j] = \int_{t[i]-w/2}^{t[i]+w/2} \int_{t[j]-w/2}^{t[j]+w/2} f_{X_n,X_{n-1}}(x_n, x_{n-1}) dx_n dx_{n-1}$$

Before proceeding, we note that each bin probability is lower than in the one-dimensional case. For example, the probability of the pair $(1.3, 1.4)$ is clearly be

lower than either just 1.3 and 1.4. We previously experienced erratic behavior in low probability regions, which drove down the rate. Additionally since we now consider minimizing over $|\mathcal{Y}|^2$ bins, rather than $|\mathcal{Y}|$ (where $|\mathcal{Y}|$ is the size of alphabet), we expect a lower minimum. We can forecast the rate will be lower when compensating pairs, satisfying the intuition that greater security must come at a cost.

Continuing just as before our basic requirement is

$$\begin{aligned} P_{Z_n,Z_{n-1}}[i,j] &= P_{X_n,X_{n-1}}[i,j] \;\forall\, i,j \\ P_{Z_n,Z_{n-1}}[i,j] &\triangleq \lambda_2 P_{S_n,S_{n-1}}[i,j] + (1-\lambda_2) P_{C_n,C_{n-1}}[i,j] \end{aligned} \tag{5.3}$$

Where λ_2 is the proportion of *pairs* used for hiding. The total number of pairs in a Markov chain of length N is $N-1$. Hiding is typically done in individual values, rather than pairs. If $\lambda_2(N-1)$ pairs are reserved for embedding, then $\lambda_2(N-1)$ coefficients available for hiding, and $(1-\lambda_2)(N-1)+1$ coefficients are available for compensating. We have the same constraints on $P_{C_n,C_{n-1}}[i,j]$ as before, as well as the problems in low probability regions, so we have a limit on the number of pairs we can use while still compensating the joint histogram:

$$\lambda_2^* \triangleq \min_{i,j \in \mathcal{H}_2} \frac{P_{X_n,X_{n-1}}[i,j]}{P_{S_n,S_{n-1}}[i,j]} \forall i,j$$

Here $\mathcal{H}_2$ is a two-dimensional region of the joint statistics. $\mathcal{H}_2 \triangleq \{i,j \,:\, P_{X_n,X_{n-1}}[i,j] \geq T'\}$, where T' is a probability threshold. Despite the

different formulation, we note that the region defined by a probability threshold is functionally equivalent to our one-dimensional region using a value threshold. In other words, for symmetric, unimodal P_X, there exists a probability threshold T' such that $\{i : P_X[i] \geq T'\} = \{i \in [T, T]\}$ for any value threshold T.

The derivation of the distribution of λ_2^* is identical to λ^* above, though with more bins to consider. The cumulative distribution of $\Gamma[i,j] \triangleq \frac{P_{X_n,X_{n-1}}[i,j]}{P_{S_n,S_{n-1}}[i,j]}$ is

$$F_{\Gamma[i,j]}(\gamma) = \sum_{k=0}^{N} \sum_{l=0}^{\lfloor \gamma k \rfloor} P_{V_S[i,j]}(k) P_{V_X[i,j]}(l)$$

and the cumulative distribution of λ_2^* is

$$F_{\lambda_2^*}(\gamma) = 1 - \left\{ \prod_{i,j} \left[1 - F_{\Gamma[i,j]}(\gamma) \right] \right\}$$

The general factors affecting data rate are the same as in one-dimensional case: cover and stego pdfs, the number of samples, and the bin width.

5.4.2 Comparing Rates of Zero K-L and Zero Matrix Divergence QIM

We now compare the 90%-safe rates of hiding for one- and two-dimensional histogram compensation to judge the cost of increased security. We experiment on dithered QIM hiding in a zero-mean bivariate Gaussian:

$$f_{X_n,X_{n-1}}(x_n, x_{n-1}) = \frac{1}{2\pi |\Sigma_X|^{1/2}} \exp - \left\{ \frac{1}{2} [x_n \; x_{n-1}] \Sigma_X^{-1} [x_n \; x_{n-1}]^T \right\}$$

where

$$\Sigma_X = \sigma_x^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

with ρ denoting the correlation coefficient. For the joint pdf of stego after dithered QIM hiding, we note that both the dither sequence and the message data sequences are independent, therefore the extension to two dimensions is trivial:

$$\begin{aligned} f_{S_n,S_{n-1}}(s_n, s_{n-1}) &= \frac{1}{\Delta} \int_{s_n-\Delta/2}^{s_n+\Delta/2} \int_{s_{n-1}-\Delta/2}^{s_{n-1}+\Delta/2} f_{X_n,X_{n-1}}(x_n, x_{n-1}) dx_n dx_{n-1} \\ &\equiv \left[f_{X_n,X_{n-1}}(x_n, x_{n-1}) * \frac{1}{\Delta} \sqcap_2 (x_n/\Delta, x_{n-1}/\Delta) \right] (s_n, s_{n-1}) \end{aligned}$$

where $*$ represents a two-dimensional convolution, and $\sqcap_2$ is a two-dimensional rectangle function. This is similar to spread spectrum hiding, except the convolution is not with a Gaussian. As with SS, we expect sharply peaked distributions to be changed more by the smoothing. When considering a two-dimensional histogram, both the variance σ_X^2 and ρ affect the concentration of probability. As with the one-dimensional case, a low variance is associated with a large peak. Additionally, as noted in Chapter 4, a strongly correlated distribution, corresponding to a large ρ value, has probability concentrated on the center line; this concentration is be spread by hiding, just as in the SS case.

We calculated the 90%-safe rates for both one- and two- dimensional histogram (empirical matrix) restoration. In Figure 5.11, we see that for data that is weakly or not correlated, there is a large cost to compensating the empirical matrix. Ad-

ditionally the security gain in this cases is quite small, as the divergence of the histogram is nearly equal to the divergence of the empirical matrix. For transform coefficients, which are only weakly correlated, it is probably not worthwhile to consider joint statistics. However we are surprised to find that for strongly correlated covers, the matrix compensation rate approaches and even passes the histogram compensation rate. We note that since

$$P_Z[i] = \sum_j P_{Z_n,Z_{n-1}}[i,j]$$

then if $P_{Z_n,Z_{n-1}} = P_{X_n,X_{n-1}}$, then $P_Z = P_X$. Therefore a perfect matrix restoration implies perfect histogram restoration, and a higher rate can be achieved for greater security! Though this seems to violate intuition, we point out that the ratio of distributions is non-linear, and it is difficult to guess results beforehand. In particular, although these ratios are equivalent

$$\frac{\sum_j P_{X_n,X_{n-1}}[i,j]}{\sum_j P_{S_n,S_{n-1}}[i,j]} = \frac{P_X[i]}{P_S[i]}$$

these are not

$$\sum_j \left(\frac{P_{X_n,X_{n-1}}[i,j]}{P_{S_n,S_{n-1}}[i,j]} \right) \neq \frac{P_X[i]}{P_S[i]}.$$

and we do not expect any obvious relationship between the distribution of $\Gamma[i]$ and $\Gamma[i,j]$.

There is a key caveat to this finding however. For the matrix restoration, we have used a probability threshold to decide which bins to hide in. Though

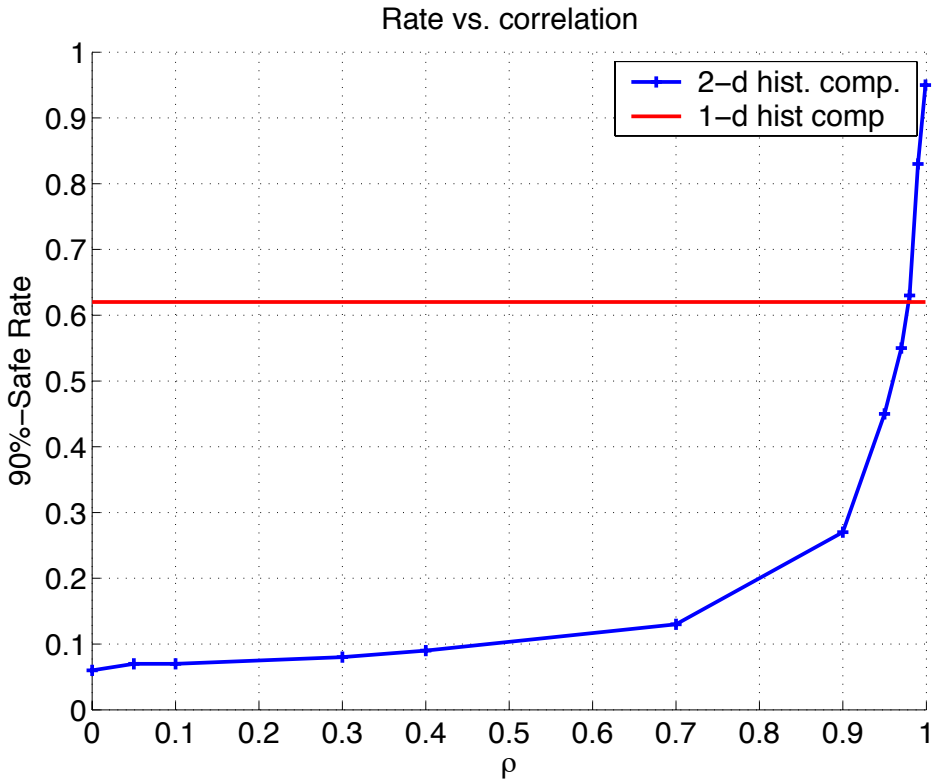

Figure 5.11: A comparison of the rates guaranteeing perfect marginal and joint histogram restoration 90% of the time. Correlation does not affect the marginal statistics, so the rate is constant. All factors other than ρ are held constant: $N = 10000$, $w = 0.1$, $\sigma_X = 1$, $\Delta = 2$. Surprisingly, compensating the joint histogram can achieve higher rates than the marginal histogram.

the probability threshold can be achieved in the one dimensional case by a value threshold, the opposite is not true. That is, a value threshold is not equivalent to a probability threshold in the joint distribution. We find that if instead a value threshold is used a for matrix restoration, it is not possible to achieve these rates. Geometrically, the probability threshold corresponds to an ellipse-

like region in the joint distribution, whereas the value threshold is a square region. The former is better fitted to the distribution of correlated data. In particular, the balance between the number of embedding coefficients and λ^* is more difficult to do with a square region. The smallest possible change in the threshold causes a large change in both $G(T)$ and λ^*. Therefore using a square region (i.e. a value threshold) reduces the rate significantly. In the one-dimensional case we had practical difficulties implementing the value threshold, and our solution achieved a lower rate than the ideal rate. We expect further difficulties with a probability threshold, and in practice these rates are may be difficult to achieve.

5.5 Summary

We have analyzed a hiding scheme designed to increase security by explicit restoration of the pdf. We believe there are two advantages of this approach over earlier compensation schemes. First, continuous valued covers, for example, uncompressed transform coefficients, can be used. Second, because a proven QIM method is used for hiding, the scheme is robust against some noise, attacks, and image processing. By first considering allowing a small amount of divergence between cover and stego by not compensating a small low-probability region, we find the tradeoff using compensation is better than the tradeoff achieved by simply

embedding less. We continue the analysis be examining perfect compensation. We derive expressions to evaluate the rate guaranteeing perfect security within a specified probability for both the ideal case and a practical implementation. We expand these results to the case of protecting against steganalysis using a level of dependency. By comparing these results with the simpler case we find that, theoretically, empirical matrix compensation can outperform histogram compensation. However, this only occurs in very strongly correlated data, and caution that it may be very difficult to achieve these results in practice.

Chapter 6

Future Work and Conclusions

The proliferation of steganographic tools has created a demand for powerful means to detect hidden data. The primary focus of this dissertation is to employ a systematic approach to the study of steganalysis allowing both the derivation of optimal bounds and the design of practical solutions. Using a detection-theoretic approach, we investigated the inherent detectability of several commonly used data hiding techniques, devised methods to detect these schemes, and used this knowledge to devise and analyze a means of escaping detection.

Though our approach has gained satisfying results in the study of steganography, we acknowledge there are problems yet to be solved. We conclude with a look to future research directions which we believe will advance the study of stealthy transmission of, and interception of, hidden data in images.

6.1 Improving Model of Images

We have seen that using a Markov chain image model has increased the accuracy of our steganalysis results, by allowing one level of dependency. Markov random field (MRF) models, with larger neighborhoods of dependency, have had some success in characterizing images. Many other models [83, 80] for describing spatial and transform domain statistical characteristics have been proposed. However, while a more complex model may be a more accurate representation of image data, a problem with increasing the complexity of the model is the increasing inaccuracy of estimates of the model for a given number of samples [15].

Typically statistical models assume the distribution is the same for all image pixels, that is, the random process is stationary. However we have seen some data hiding methods [81, 37, 31] that adaptively embed data according to the local statistics, suggesting that a careful steganalyst should consider a variation in statistics across the image. A piece-wise stationarity assumption, in which a different stationary random process is assumed for several regions within the image, is perhaps a better way to model real images while remaining analytically tractable. There is however a significant challenge with this approach: how to choose regions that accurately group statistically related pixels. Certainly a

smaller region is more likely to follow a stationarity assumption, however a smaller region has fewer samples with which to estimate the distribution of the region. Grouping pixels from the same distribution is essentially image segmentation, a very difficult problem and an active field of research. By leveraging the results of this field, an appropriate means of selection may be found.

In addition to variation within an image, we have observed in our testing a noticeable disparity in detector effectiveness according to the source of the cover image. By source we mean both the scene represented in the image and the technology used to create a digital representation of the scene. For example, small sections of a very high resolution image are much more homogenous than a lower resolution image of the same size. Assumptions based on stationarity and strong correlation therefore work better for the former than the latter. We have attempted to account for this in our experiments by using images from several different sources; however the choice of an "ideal" natural image database remains an ill-posed problem. To improve detection on images with widely varying sources, for example images grabbed randomly off of the Internet, it may be beneficial to use a two-stage classifier. The first stage classifies the image into rough groups based on source, the second stage discriminates between cover and stego within these groups.

6.2 Accurate Characterization of Non-Optimal Detection

A key advantage of the detection-theoretic approach is provable bounds on detector performance giving a steganographer a guarantee of security. A decrease in hiding rate is generally accepted to effect an increase in security. However, anticipating optimal detection may be too pessimistic, causing a greater sacrifice of rate than is actually necessary to avoid detection. Practical detectors will always fall short of optimal, due to the insufficiency of information practically available. In particular, the steganalyst's lack of knowledge of the cover distribution means she or he has to use general observations on natural images, either directly or through supervised learning. If the exact cost to detector performance caused by estimating this information could be accurately characterized, the steganographer could hide at an increased rate while still remaining effectively stealthy. Additionally, the steganalyst will know when the practical limits of detection have been reached. Accurately characterizing the difference between optimal detection and the best that can be realistically done is a very challenging problem, but the potential benefits are great.

6.3 Summary

In this thesis we implemented a detection-theoretic approach to the analysis of steganographic security from both the detector's and hider's point of view. This theory is well-developed and is naturally suited to the steganalysis problem. This approach allows us to estimate the performance of optimal statistical detection of hiding in images.

In addition to characterizing optimal steganalysis under idealized conditions, we developed methods for practical detection in realistic scenarios. Specifically, we develop tools to detect three general classes of data hiding in natural images: least significant bit (LSB), quantization index modulation (QIM), and spread spectrum (SS).

Though powerful detection schemes exploiting image dependencies exist, systematic approaches to steganalysis have typically focused on an independent and identically distributed (i.i.d.) assumption. We extend the detection-theoretic approach to the next logical step by using a Markov chain model of hiding media, thus allowing a systematic approach using a measure of dependency.

Finally, we leveraged our steganalysis knowledge to design a system to evade optimal steganalysis. In addition to designing a system which successfully reduces the effectiveness of previously successful detection for dithered QIM, this analysis

is also used to derive a formulation of the rate of secure hiding for arbitrary cover distributions.

Steganalysis and steganography are complex problems, and there are many avenues available for further exploration. Image data is difficult to succinctly characterize. We note that a more complex image model could be used, allowing for more accurate statistical description of images. Additionally, the detection theory provides estimates of the performance of optimal steganalysis. However optimal steganalysis is practically impossible. If the deviation from optimality could be accurately characterized, an estimate of inherent practical detectability may be possible. Finally we have generally focused on grayscale still images. However the methods we presented here can be applied to the study of data hiding in color images, video, and audio.

Bibliography

[1] V. Anantharam. A large deviations approach to error exponents in source coding and hypothesis testing. *IEEE Trans. on Information Theory*, 36(4):938–943, 1990.

[2] T. Aura. Practical invisibility in digital communication. *Lecture Notes in Computer Science: 1st Int'l Workshop on Information Hiding*, 1174:265–278, 1996.

[3] I. Avcibas, N. Memon, and B. Sankur. Steganalysis using image quality metrics. In *Proc. IST/SPIE's 13th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2001.

[4] I. Avcibas, N. Memon, and B. Sankur. Image steganalysis with binary similarity measures. In *Proceedings of ICIP*, 2002.

[5] I. Avcibas, N. Memon, and B. Sankur. Steganalysis using image quality metrics. *IEEE Trans. on Image Processing*, 12(2):221–229, 2003.

[6] R. Blahut. *Principles and practice of information theory.* Addison Wesley, 1987.

[7] R. Bohme and A. Westfeld. Breaking Cauchy model-based JPEG steganography with first-order statistics. In *Proc. 9th ESORICS*, France, Sep 2004.

[8] R. Bohme and A. Westfeld. Exploiting preserved statistics for steganalysis. In *Proc. 6th Int'l Workshop on Information Hiding*, 2004.

[9] A. Brown. S-tools 4.0, http://www.snapfiles.com/get/stools.html.

[10] C. Cachin. An information theoretic model for steganography. *Lecture Notes in Computer Science: 2nd Int'l Workshop on Information Hiding*, 1525:306–318, 1998.

[11] M. U. Celik, G. Sharma, and A. Tekalp. Universal image steganalysis using rate-distortion curves. In *Proc. IST/SPIE's 16th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan 2004.

[12] R. Chandramouli, M. Kharrazi, and N. Memon. Image steganography and steganalysis: Concepts and practices. In *Proceedings of Second Int'l Workshop on Digital Watermarking*, pages 35–49, 2003.

[13] R. Chandramouli and N. Memon. Analysis of LSB based image steganography techniques. In *Proceedings of ICIP*, pages 1019–1022, 2001.

[14] B. Chen and G. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Info. Theory*, 47(4):1423–1443, May 2001.

[15] V. Cherkassky and F. Mulier. *Learning From Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., 1988.

[16] K. L. Chung. *Markov Chains with stationary transition probabilities*. Spring-Verlag, 1960.

[17] M. Costa. Writing on dirty paper. *IEEE Trans. Info. Theory*, IT-29(3):439–441, May 1983.

[18] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[19] I. Cox, J. Kilian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. on Image Processing*, 6(12):1673–1687, 1997.

[20] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath. Detection of hiding in the least significant bit. In *Proceedings of Conference on Information Sciences and Systems (CISS)*, 2003.

[21] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath. Detection of hiding in the least significant bit. *IEEE Trans. on Signal Processing, Supplement on Secure Media I*, 52(10):3046–3058, 2004.

[22] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 2 edition, 1998.

[23] S. Draper, P. Ishwar, D. Molnar, V. Prabhakaran, K. Ramchandran, D. Schonberg, and D. Wagner. An analysis of empirical PMF based tests for least significant bit image steganography. In *Proc. 7th Int'l Workshop on Information Hiding*, 2005.

[24] S. Dumitrescu, X. Wu, and Z. Wang. Detection of LSB steganography via sample pair analysis. *IEEE Trans. on Signal Processing*, 51(7):1995–2007, 2003.

[25] J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod. Scalar Costa scheme for information embedding. *IEEE Trans. on Signal Processing*, 51(4):1003–1019, 2003.

[26] J. J. Eggers, R. Bauml, and B. Girod. A communications approach to image steganography. In *Proc. IST/SPIE's 14th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2002.

[27] FortKnox, software: http://www.clickok.co.uk/steg/index.html.

[28] H. Farid. Detecting stenographic messages in digital images. Technical report, Dartmouth College, Computer Science, 2001.

[29] E. Franz. Steganography preserving statistical properties. In *5th International Working Conference on Communication and Multimedia Security*, 2002.

[30] J. Fridrich. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes. In *Proc. of Sixth Information Hiding Workshop*, May 2004.

[31] J. Fridrich and M. Goljan. Digital image steganography using stochastic modulation. In *Proc. IST/SPIE's 15th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan 2003.

[32] J. Fridrich and M. Goljan. On estimation of secret message length in LSB steganography in spatial domain. In *Proc. IST/SPIE's 16th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2004.

[33] J. Fridrich, M. Goljan, and R. Du. Reliable detection of LSB steganography in color and grayscale images. In *Proc. ACM Workshop on Multimedia Security*, Ottawa, Canada, 2001.

[34] J. Fridrich, M. Goljan, and D. Hogea. Attacking the OutGuess. In *Proceedings of ACM Workshop on Multimedia and Security*, Juan-Pins, France, Dec 2002.

[35] J. Fridrich, M. Goljan, and D. Hogea. Steganalysis of JPEG images: Breaking the F5 algorithm. In *Lecture notes in computer science: 5th International Workshop on Information Hiding*, volume 2578, pages 310–323, 2002.

[36] J. Fridrich, M. Goljan, and D. Hogea. New methodology for breaking steganographic techniques for JPEGs. In *Proc. IST/SPIE's 15th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan 2003.

[37] J. Fridrich, M. Goljan, P. Lisonek, and D. Soukal. Writing on wet paper. In *Proc. IST/SPIE's 17th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2005.

[38] J. Fridrich, M. Goljan, and D. Soukal. Perturbed quantization steganography with wet paper codes. In *Proc. of ACM Multimedia and Security Workshop*, Sep 2004.

[39] S. I. Gel'Fand and M. S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, Jan. 1979.

[40] A. Gersho and R. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.

[41] R. C. Gonzalez and R. Woods. *Digital image processing*. Addison Wesley, 1992.

[42] A. Goteti and P. Moulin. QIM watermarking games. In *Proceedings of ICIP*, Singapore, Oct 2004.

[43] P. Guillon, T. Furon, and P. Duhamel. Applied public-key steganography. In *Proc. IST/SPIE's 14th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2002.

[44] A. Habibi. Comparison of n-th order DPCM encoder with linear transformations and block quantization techniques. *IEEE Trans. on Communication Technology*, COM-19(6):948–956, Dec 1971.

[45] J. J. Harmsen and W. A. Pearlman. Steganalysis of additive noise modelable information hiding. In *Proc. IST/SPIE's 15th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2003.

[46] M. T. Hogan, N. J. Hurley, G. C. M. Silvestre, F. Balado, and K. M. Whelan. ML detection of steganography. In *Proc. IST/SPIE's 17th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2005.

[47] InThePicture, software: http://www.intar.com/ITP/itpinfo.htm/.

[48] Invisible Secrets, software: http://www.invisiblesecrets.com/.

[49] JSTEG, software: http://www.theargon.com/archives/steganography/.

[50] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.

[51] T. Joachims. Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[52] N. F. Johnson, Z. Duric, and S. Jajodia. *Information Hiding: Steganography and Watermarking - Attacks and Countermeasures*. Kluwer Academic Publishers, 2001.

[53] M. Kharrazi, H. T. Sencar, and N. Memon. Benchmarking steganographic and steganalysis techniques. In *Proc. IST/SPIE's 17th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2005.

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

[55] E. T. Lin and E. J. Delp. A review of data of hiding in digital images. In *Proceedings of the Image Processing, Image Quality, Image Capture Systems Conference (PICS '99)*, pages 274–278, Savannah, Georgia, 1999.

[56] S. Lyu and H. Farid. Detecting hidden messages using higher-order statistics and support vector machines. In *Lecture notes in computer science: 5th International Workshop on Information Hiding*, volume 2578, 2002.

[57] S. Lyu and H. Farid. Steganalysis using color wavelet statistics and one-class support vector machines. In *Proc. IST/SPIE's 16th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2004.

[58] B. Macq, J. Dittmann, and E. J. Delp. Benchmarking of image watermarking algorithms for digital rights managment. *Proc. of the IEEE*, 92(6):971–983, 2004.

[59] A. Martin, G. Sapiro, and G. Seroussi. Is image steganography natural? HP labs tech report HPL-2004-39, 7 March 2004.

[60] L. Marvel, C. G. Boncelet Jr., and C. T. Retter. Spread spectrum image steganography. *IEEE Trans. on Image Processing*, 8(8):1075–1083, 1999.

[61] M. K. Mihçak and R. Venkatesan. Blind image watermarking via derivation and quantization of robust semi-global statistics. In *Proceedings of ICASSP*, May 2002.

[62] M. K. Mihçak, R. Venkatesan, and M. Kesal. Cryptanalysis of discrete-sequence spread spectrum watermarks. *Lecture Notes in Computer Science: 5th Int'l Workshop on Information Hiding*, 2578:226–246, 2003.

[63] P. Moulin and Y. Wang. New results on steganographic capacity. In *Proceedings of Conference on Information Sciences and Systems (CISS)*, 2004.

[64] S. Natarajan. Large deviations, hypothesis testing, and source coding for finite Markov chains. *IEEE Trans. on Information Theory*, 31(3):360–365, 1985.

[65] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[66] R. L. Pickholtz, D. L. Schilling, and L. B. Milstein. Theory of spread-spectrum communications – a tutorial. *IEEE Trans. on Communications*, 30(5):855–884, 1982.

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

[68] N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, Washington DC, 2001.

[69] N. Provos and P. Honeyman. Detecting steganographic content on the internet. In *ISOC NDSS'02*, San Diego, CA, 2002.

[70] A. Rangarajan and R. Chellappa. Markov random field models in image processing. In *The handbook of brain theory and neural networks*, pages 564–567, 1995.

[71] C. R. Rao. *Linear Statistical Inference and Its Applications*. John Wiley and Sons, 1965.

[72] B. Roue, P. Bas, and J.-M. Chassery. Improving LSB steganalysis using marginal and joint probabilistic distributions. In *Proc. of ACM Multimedia and Security Workshop*, pages 75–80, Sept 2004.

[73] J. J. K. O. Ruanaidh and T. Pun. Rotation, scale and translation invariant digital image watermarking. In *Proceedings of ICIP*, pages 536–539, Santa Barbara, CA, 1997.

[74] StegoArchive.com: http://www.stegoarchive.com.

[75] P. Sallee. Model-based steganography. In *Proceedings of Second Int'l Workshop on Digital Watermarking*, pages 154–167, 2003.

[76] P. Sallee. Model-based methods for steganography and steganalysis. *International Journal of Image and Graphics (IJIG)*, 5(1):167–190, 2005.

[77] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–10, 1979.

[78] M. Sidorov. Hidden Markov models and steganalysis. In *Proc. of ACM Multimedia and Security Workshop*, pages 63–67, Sept 2004.

[79] M. Sidorov. A statistical steganalysis for digital images. In *Proc. of the International I and S Workshop*, pages 34–36, Jan 2004.

[80] E. Simoncelli. Statistical models for images: Compression restoration and synthesis. In *Proceedings of 31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 1997.

[81] K. Solanki, N. Jacobsen, U. Madhow, B. S. Manjunath, and S. Chandrasekaran. Robust image-adaptive data hiding based on erasure and error correction. *IEEE Transactions on Image Processing*, 13(12):1627–1639, Dec 2004.

[82] K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran. Statistical restoration for robust and secure steganography. To appear *ICIP 2005*.

[83] A. Srivastava, A. Lee, E. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18:17–33, 2003.

[84] K. Sullivan, Z. Bi, U. Madhow, S. Chandrasekaran, and B. S. Manjunath. Steganalysis of quantization index modulation data hiding. In *Proceedings of ICIP*, pages 1165–1168, Singapore, Oct 2004.

[85] K. Sullivan, O. Dabeer, U. Madhow, B. Manjunath, and S. Chandrasekaran. LLRT based detection of LSB hiding. In *Proceedings of ICIP*, volume 1, pages 497–500, Barcelona, Spain, Sep 2003.

[86] K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath. Steganalysis of spread spectrum data hiding exploiting cover memory. In *Proc. IST/SPIE's 17th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan 2005.

[87] K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran. Steganalysis of Markov cover data with applications to images. Submitted to *IEEE Trans. on Information Forensics and Security*.

[88] R. Tzschoppe, R. Bauml, J. B. Huber, and A. Kaup. Steganographic system based on higher-order statistics. In *Proc. IST/SPIE's 15th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Jan 2003.

[89] R. van Schyndel, A. Tirkel, and C. Osborne. A digital watermark. In *Proceedings of ICIP*, volume 2, pages 86–90, Austin, TX, 1994.

[90] R. Venkatesan and M. H. Jakubowski. Image watermarking with better resilience. In *Proceedings of ICIP*, Vancouver, British Columbia, Canada, 2000.

[91] R. Venkatesan, V. Vazirani, and S. Sinha. A graph theoretic approach to software watermarking. *Lecture Notes in Computer Science: 4th Int'l Workshop on Information Hiding*, 2137:157–168, 2001.

[92] M. Vetterli and C. Herley. Wavelets and filter banks: Theory and design. *IEEE Transactions on Signal Processing*, 40(9):2207–2232, 1992.

[93] R. F. Walker, P. Jackway, and I. Longstaff. Improving co-occurrence matrix feature discrimination. In *Proc. of Digital Image Computing: Techniques and Applications (DICTA)*, pages 643–648, Dec 1995.

[94] G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[95] M. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64, Feb 1997.

[96] M. Wand and M. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

[97] Y. Wang and P. Moulin. Steganalysis of block-DCT steganography. In *Proc. IEEE Workshop on Statistical Signal Processing*, St Louis, MO, Sep 2003.

[98] Y. Wang and P. Moulin. Steganalysis of block-structured stegotext. In *Proc. IST/SPIE's 16th Annual Symposium on Electronic Imaging Science and Technology*, San Jose, CA, 2004.

[99] A. Westfeld. High capacity despite better steganalysis (F5 - a steganographic algorithm). In *Lecture notes in computer science: 4th International Workshop on Information Hiding*, volume 2137, pages 289–302, 2001.

[100] A. Westfeld and A. Pfitzmann. Attacks on steganographic systems. In *Lecture notes in computer science: 3rd International Workshop on Information Hiding*, 1999.

[101] R. Wolfgang and E. Delp. A watermark for digital images. In *Proceedings of ICIP*, pages 219–222, Lausanne, Switzerland, 1996.

[102] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp. The effect of matching watermark and compression transforms in compressed color images. In *Proceedings of ICIP*, Chicago, Illinois, Oct. 1998.

[103] J. J. Yu, J. W. Han, S. C. O, S. Lee, and I. H. Park. A secure steganographic scheme against statistical analyses. In *Lecture Notes in Computer Science: Proceedings of Second Int'l Workshop on Digital Watermarking*, volume 2939, pages 497–507, 2004.

[104] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. on Information Theory*, 48(6):1250–1276, 2002.

[105] O. Zeitouni, J. Ziv, and N. Merhav. When is the generalized likelihood ratio test optimal. *IEEE Transactions on Information Theory*, 38(5):1597–1602, 1992.

[106] T. Zhang and X. Ping. A new approach to reliable detection of LSB steganography in natural images. *Signal Processing*, 83(10):2085–2093, 2003.

[107] J. Zolner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. Wicke, and G. Wolf. Modeling the security of steganographic systems. In *Lecture notes in computer science: 2nd International Workshop on Information Hiding*, volume 1525, pages 345–355, 1998.

Appendix A

Glossary of Symbols and Acronyms

List of Symbols

n, m, i, j, k various indexes, n is typically a sample index, i, j a value or bin index

A acceptance region of detector, page 38

$\mathbf{A}$ regularity constraint matrix, page 52

α probability of false alarm, or SS scaling parameter page 38 or 91

B hidden message bits, page 43

Bl blockiness measure, page 89

C output of standard quantizer or compensating coefficients, page 65 or 127

^{c} denotes complement of a set, general use

$\mathfrak{C}$ compensated region, page 129

D dither value for dithered QIM, or SS message bearing signal, page 67 or page 90

$D(\cdot||\cdot)$ Kullback-Leibler divergence (relative entropy), page 41

$D(\mathbf{M}^{(X)}, \mathbf{M}^{(S)})$ matrix divergence between Markov chains X and S, page 85

$\delta(\cdot)$ generic detector, page 38

Δ^* step size for uniform scaler quantization, page 66

Δ step size for uniform scaler QIM, page 63

η_{ij} number of transitions from i to j in Markov chain, page 83

ϵ bound on Kullback-Leibler divergence, page 30

$f_X(x), f_S(s)$ probability density function (pdf) of cover and stego, page 126

$F_X(x)$ cumulative distribution function of random variable X: $P(X \leq x)$, general use

$g(\cdot)$ scaling function for practical threshold, page 145

$G(\cdot)$ number of embeddable coefficients as a function of threshold, page 143

Γ random variable, ratio of cover histogram to stego histogram, page 134

γ value of Γ, page 134

$H(\cdot)$ entropy of random variable, general use

$H_b(\cdot)$ entropy of binary random variable, page 27

H_0, H_1 cover, stego hypotheses, page 37

$\mathcal{H}$ hiding region for zero divergence statistical compensation, page 142

I number of bins, page 126

K composite hypothesis of stego hiding at any rate, page 45

κ ratio of matrix divergence to K-L divergence, page 86

$L(\cdot)$ log-likelihood statistic, page 39

$L(\cdot)_{\chi^2}$ chi-squared statistic, page 48

$L_{\mathbf{approx.}}(\cdot)$ approximate log-likelihood statistic, page 51

λ percentage of coefficients used for hiding, page 127

λ_2 percentage of pairs used for hiding, joint compensation case, page 151

λ^* percentage of coefficients available for hiding, perfect compensation, page 129

λ_2^* same as λ^* for joint compensation case, page 151

$\mathbf{M}$ empirical matrix of Markov chain, page 85

$\mathbf{M}^{(X)}, \mathbf{M}^{(S)}$ cover, stego empirical matrices, page 85

N number of samples, page 37

P_X, P_S function form of PMFs of cover, stego, page 65

$P(E)$ probability of event E, general use

$\mathbf{p}^{(X)}, \mathbf{p}^{(S)}$ vector form of PMF cover and stego, page 40

$\widehat{\mathbf{p}}^{(X)}$ estimate of cover PMF, page 51

q_b quantizer indexed by message, page 62

$\mathbf{q}$ estimated PMF of received data (normalized histogram), page 40

$\mathbf{Q}_R$ matrix corresponding to linear transformation of LSB hiding at rate R, page 44

$Q(\cdot)$ complementary Gaussian function, page 47

R hiding rate, page 43

$\mathbb{R}$ real line

$\sqcap(t)$ rectangular function : $u(t+1/2) - u(t-1/2)$, page 138

S, s stego random variable, value of stego random variable, page 37

T threshold for statistical compensation, page 142

$\mathbf{T}$ Markov chain transition matrix, page 83

σ_X standard deviation of random variable X

τ threshold for hypothesis test, page 38

w bin width of histogram, page 64

X, x cover random variable, value of cover random variable, page 37

$\mathcal{X}^*$ quantization range (Voronoi region) for uniform scalar quantization, page 65

$\mathcal{X}$ quantization range (Voronoi region) for uniform scalar QIM, page 66

χ^2 closeness-of-fit statistic, used for LSB detection, page 15

$\{Y_n\}_{n=1}^N$ received data samples, page 37

$\mathbf{y}$ received data vector, page 40

$\mathcal{Y}$ alphabet of data, page 38

Z statistically compensated stego, page 127

$\mathbf{Z}_n$ indicator vector, page 46

List of Acronyms

AWGN additive white Gaussian noise

bpp bits per pixel

bpnz-DCT bits per non-zero DCT coefficient

CACD Canon digital camera

CPCD Corel photo CD

COM center of mass

DCT discrete cosine transform

DFT discrete Fourier transform

DWT discrete wavelet transform

DOQQ digital orthophoto quarter-quadrangle (Set of aerial images.)

HCF histogram characteristic function

MISE mean integrated squared error

HMM hidden Markov model

HPDM histogram-preserving data-mapping

HVS human visual system

IQM image quality metrics

i.i.d. independent and identically distributed

JPEG joint photographic experts group (Image compression schemes.)

K-L Kullback-Leibler

KLT Karhunen-Loeve transform

LSB least significant bit

LRT likelihood ratio test

LLRT log likelihood ratio test

MC Markov chain

MCR message to cover (power) ratio

MRF Markov random field

MSE mean squared error

MMSE minimum mean squared error

pdf probability density function

PNG portable network graphics (Image compression scheme.)

PMF probability mass function

PR psuedorandom

PSNR peak signal to noise ratio

QIM quantization index modulation

ROC receiver operating characteristics

RS regular/singular (Used to denote detection method using sets with these names.)

SS spread spectrum

SSIS spread spectrum image steganography

SVM support vector machine