Benchmarking for Steganography

Conference Paper · October 2008

DOI: 10.1007/978-3-540-88961-8_18 · Source: DBLP

2 authors:

Tomás Pevný
Czech Technical University in Prague
82 PUBLICATIONS **3,982** CITATIONS
SEE PROFILE

Jessica Fridrich
Binghamton University
129 PUBLICATIONS **12,569** CITATIONS
SEE PROFILE

Benchmarking for Steganography

Tomáš Pevný[1] and Jessica Fridrich[2]

[1] Department of CS, SUNY Binghamton, pevnak@gmail.com
[2] Department of ECE, SUNY Binghamton, fridrich@binghamton.edu

Abstract. With the increasing number of new steganographic algorithms as well as methods for detecting them, the issue of comparing security of steganographic schemes in a fair manner is of the most importance. A fair benchmark for steganography should only be dependent on the model chosen to represent cover and stego objects. In particular, it should be independent of any specific steganalytic technique. We first discuss the implications of this requirement and then investigate the use of two quantities for benchmarking—the KL divergence between the empirical probability distribution of cover and stego images and the recently proposed two-sample statistics called Maximum Mean Discrepancy (MMD). While the KL divergence is preferable for benchmarking because it is the more fundamental quantity, we point out some practical difficulties of computing it from data obtained from a test database of images. The MMD is well understood theoretically and numerically stable even in high-dimensional spaces, which makes it an excellent candidate for benchmarking in steganography. We demonstrate the benchmark based on MMD on specific steganographic algorithms for the JPEG format.

1 Introduction

Up until now, the security of steganographic systems was compared by reporting detection results for a specific blind steganalyzer [16, 24, 6]. This is clearly undesirable because the comparison is dependent on the steganalyzer feature set, the machine learning engine (SVM, neural network, etc.), and a functional assigning a single numerical value to the ROC curve (total minimal decision error [16, 38], probability of detection for fixed false alarm rate [26], or false alarm for probability of detection 50% [20], accuracy [13], etc.). The goal of this paper is to provide a practical method for comparing security of steganographic systems that is free from such arbitrary choices and thus provides a more fundamental measure of security than previously proposed measures. We interpret the selection of the feature set as a low-dimensional *model* of covers and compute the steganographic security directly in the model space from empirical data obtained from a database of cover and stego images. We consider two different measures, each one of which has its own advantages and disadvantages–the Kullback-Leibler divergence and the Maximum Mean Discrepancy (MMD) two-sample statistics.

In the next section, we introduce some basic concepts and explain the motivation for our approach. In Section 3, we describe a method for benchmarking

steganographic systems using the KL divergence and discuss its limitations. The MMD is introduced in Section 4. We use MMD to compute benchmark values for selected known steganographic algorithms in Section 5. In the same section, we discuss the experiments and compare the benchmark to results obtained using SVMs. The paper is concluded in Section 6.

2 Steganographic security and cover models

Denoting $\mathcal{C}$ the set of all covers c, Cachin's definition of steganographic security is based on the assumption that the selection of covers from $\mathcal{C}$ can be described by a random variable c on $\mathcal{C}$ with probability distribution function (pdf) P. A steganographic scheme, S, is a mapping $\mathcal{C} \times \mathcal{M} \times \mathcal{K} \to \mathcal{C}$ that assigns a new (stego) object, $s \in \mathcal{C}$, to each triple (c, M, K), where $M \in \mathcal{M}$ is a secret message selected from the set of communicable messages, $\mathcal{M}$, and $K \in \mathcal{K}$ is the steganographic secret key. Assuming the covers are selected with pdf P and embedded with a message and secret key both randomly (uniformly) chosen from their corresponding sets, the set of all stego images is again a random variable s on $\mathcal{C}$ with pdf Q. The measure of statistical detectability is the Kullback–Leibler divergence [5]

$$D(P||Q) = \sum_{c \in \mathcal{C}} P(c) \log \frac{P(c)}{Q(c)}. \tag{1}$$

When $D(P||Q) < \epsilon$, the stego system is called ϵ-secure.

The KL divergence is a very fundamental quantity because it provides bounds on the best possible detector one can build [8]. Thus, at least in theory, we could benchmark steganographic schemes by deriving Q from S and P and evaluating the KL divergence analytically. Of course, the real difficulty is that we have little information about the probability distributions involved due to the large dimensionality of the set $\mathcal{C}$. This problem is typically solved by working with simplified models of cover objects. There are basically two choices: 1) analytical models, in which c is modeled as a sequence of iid random variables [28] (or Markov chains [39]) and 2) high-dimensional models based on features extracted from the cover/stego objects [32, 13, 2, 1, 43, 36, 26, 11]. The major advantage of analytical models is that we may be able to compute the distribution of stego images and derive the relationship between the KL divergence and the amount of embedded data [7]. The weakness of this approach lies in the fact that the analytical models are too simple to capture complex cover objects, such as digital images. Often, it is not difficult to detect the "provably secure" steganographic scheme by using a better model of cover objects and designing appropriate test statistics. As an example, we cite the successful attacks [32] on steganographic systems that preserve first order statistics of DCT coefficients in JPEG images [33, 29, 19, 10, 39]. Since at this point, there are no analytically tractable models that truthfully describe natural images, we turned our attention to models based on features.

Such models far better capture the complexity of cover objects but are no longer amenable to analytical study and thus we cannot derive the KL divergence (or some other statistic) analytically. Instead, we estimate it by calculating features from a large number of covers and stego objects. This approach, however, is not without problems. First, we have the uncomfortable dependence on the database of test images, and second, it is not clear how many bits we should be embedding in each image. We postpone discussion of the message length to subsequent sections. Presumably, the issue of the size and diversity of the database can be dealt with by including sufficiently many images. While the database choice is crucial for spatial domain steganography, it is less critical for steganography in JPEG images because JPEG compression removes high-frequency details from images and thus essentially narrows down the space of covers $\mathcal{C}$. Although in this paper our covers will be digital images in the JPEG format, the methods proposed here are by no means limited to such covers and can be extended to other objects, such as raw images or audio.

3 Benchmarking steganographic schemes

In this section, we explain the basic ingredients of our benchmark and study the feasibility of using the KL divergence for benchmarking.

3.1 Model selection

As explained in Section 1, we model covers through a set of numerical features calculated from them. Formally, the model is a mapping

$$\psi : \mathcal{C} \to \mathbb{R}^d \tag{2}$$

that assigns a d-dimensional feature vector $x = \psi(c)$ to every cover. This feature vector represents the cover in $\mathbb{R}^d$ and we interpret ψ as the cover model. Consequently, the benchmark that we propose will necessarily depend on the model. The mapping ψ induces two random variables on $\mathbb{R}^d$, $\psi(\mathsf{c})$ and $\psi(\mathsf{s})$, with their corresponding pdfs p and q, reserving from now on the letter x for features from covers and y for features from stego images.

By projecting $\mathcal{C}$ onto $\mathbb{R}^d$ for some "reasonably small" d, we obviously lose a lot of information. It is important that we preserve those properties of covers that typically get disturbed by steganographic embedding. Different authors proposed different feature sets for applications in blind as well as targeted steganalysis. In this paper, we selected the 274-dimensional feature vector described in [32]. The SVM steganalyzer based on this feature can reliably detect a large number of current steganographic techniques and provides state-of-the-art results based on comparative studies reported in [24, 38, 31].

3.2 Stego images

Given two steganographic schemes, S_1 and S_2, we wish to know which method is more secure (less statistically detectable). This answer, however, will generally

depend on how the steganography is used. It is well possible that S_1 may be more detectable than S_2 for one payload size and less detectable for a different payload size. For example, methods that use matrix embedding [14] exhibit sharp non-linear decrease in detectability with decreasing payload due to significantly lower number of embedding changes, while other methods do not allow matrix embedding (e.g., adaptive schemes). Moreover, some steganographic algorithms are inherently limited to binary codes, such as methods based on perturbed quantization [25, 15], while methods that use ± 1 type of embedding can utilize more powerful ternary codes [14]. Thus, one steganographic method can be embedding significantly higher payload than some other method for the same distortion budget. Fixing the distortion budget instead of the payload would, however, benchmark the type of embedding operation rather than the whole scheme.

Perhaps, we first need to ask what it is that we want our benchmark to measure. If our goal was to evaluate the statistical detectability under conditions that somehow simulate real-life usage, we would need to know the statistical distribution of payloads that are typically embedded. It is, however, completely unclear if we can assume anything reasonable about this prior distribution. A tempting possibility is to choose an approach similar in spirit to the steganalysis benchmark proposed by Ker [22]. The reasoning is that over multiple uses of the stego channel, the relative change rate λ must converge to zero to avoid detection. Because for statistically detectable stego schemes the KL divergence is quadratic in λ, $D \approx Q\lambda^2$ as $\lambda \to 0$, it was proposed in [21] to take the constant Q for benchmarking steganalysis detectors. Adopting this approach for benchmarking steganography, we discover that the KL divergence may become non-quadratic in α due to matrix embedding. For example, for optimal codes $\lambda = H^{-1}(\alpha)$ and $D \sim \left(H^{-1}(\alpha)\right)^2$.[3] We acknowledge that this observation does not preclude the possibility to benchmark steganography in the limit $\alpha \to 0$, but do not pursue this approach further in this paper.

It seems that a reasonable option is to fix the message length with respect to the number of coefficients in the image usable for steganography. We fix the embedding rate or relative payload, α, as the ratio between the message length in bits and the number of non zero AC coefficients in the cover JPEG image (bpac). Thus, for each particular image every stego method will embed the same relative payload. By fixing $\epsilon > 0$, we could then state that a certain steganographic method becomes ϵ-secure at relative payload $\alpha(\epsilon)$. This way, the benchmark will stay compatible with the methodology accepted in previously published papers on steganalysis (see, e.g., [16]). Fixing the relative message length also makes intuitive sense because people might subconsciously use a bigger cover for large messages and a smaller cover for short messages. Also, there are some heuristic arguments that steganographic capacity might be linearly proportional to the number of pixels. Imagine that we take many pictures with a digital camera of exactly the same scene. Due to presence of random components, such as the shot noise (caused by quantum properties of light), each time we take a

[3] $H(x)$ is the binary entropy function.

picture, the pixel values will slightly vary. Subsequent in-camera processing will introduce local dependencies among the random components and thus the noise will correspond more to a Markov random field than a collection of iid variables. The entropy of this Markov field increases linearly[4] with the number of pixels [8]. Attempts to construct stego schemes around this idea include [12, 30].

3.3 KL divergence as benchmark statistics

Given a set of D database images, we generate two sets of samples

$$\mathbf{X} = (x_1, \dots, x_D),\ \mathbf{Y}(\alpha) = (y_1, \dots, y_D), \tag{3}$$

where we explicitly denoted the dependence of the samples of stego images $\mathbf{Y}$ on the relative message length α. We wish to emphasize that $x_i = \psi(c_i)$ and $y_i = \psi(s_i)$ are d-dimensional vectors (the features for cover and stego image i). Considering $\mathbf{X}$ and $\mathbf{Y}$ as vectors of D independent realizations of the random variables $\psi(\mathsf{c})$ and $\psi(\mathsf{s})$, we can estimate the KL divergence

$$D_{\mathrm{KL}}\left(\psi(\mathsf{c})\|\psi(\mathsf{s})\right) = \int\limits_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)}$$

from the empirical data (3). The high dimensionality of the feature space makes the estimation quite challenging. A practically computable benchmark cannot rely on too large a database as that would incur impractical computing requirements and storage. Realistically, we need to obtain good estimates with $10^3 - 10^5$ images. The large dimensionality eliminates most estimators of KL divergence that we can potentially use. A good review of entropy estimators is in [3]. The only estimator that can provide accurate results in high dimensional spaces is the kNN estimator [4, 37], which we now briefly describe.

3.4 The kNN estimator of KL divergence

The KL divergence can be written as

$$D_{\mathrm{KL}}(p||q) = \int\limits_{\mathbb{R}^d} p(x) \log p(x) - \int\limits_{\mathbb{R}^d} p(x) \log q(x) = -H(p) + H_{\mathsf{x}}(p, q), \tag{4}$$

where H stands for the entropy of p and $H_{\mathsf{x}}(p, q)$ for the cross-entropy. Let $\rho_k(\mathbf{X}, z)$ and $\rho_k(\mathbf{Y}, z)$ denote the radius of the smallest ball centered at $z \in \mathbb{R}^d$ that contains exactly k samples from $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then,

$$\hat{D}_{\mathrm{KL}}(p||q) = \log \frac{D}{D-1} + \frac{d}{D}\left(\sum_{i=1}^{D} \log \rho_k(\mathbf{X}, x_i) - \sum_{i=1}^{D} \log \rho_k(\mathbf{Y}, x_i)\right) \tag{5}$$

[4] Note that this argument is not in contradiction with [21] because there exist no detectors for stego schemes that use this random field for embedding.

d	2×500	2×1000	2×5000	2×10000	2×50000	2×100000	$D_{\mathrm{KL}}(p\|\|q)$
1	24.86%	23.62%	19.62%	16.05%	11.06%	9.41%	2
10	50.25%	41.15%	38.13%	38.14%	33.52%	32.58%	2
100	—	—	—	45.73%	44.10%	45.24%	2
200	—	—	—	—	45.45%	44.40%	2
300	—	—	—	—	—	50.66%	2

Table 1. Relative error of the KL-divergence estimate for two multi-variate Gaussian distributions for various combination of sample sizes, D, and data dimensionality d. The number of nearest neighbors was set to $k = \sqrt{D}$. Some combinations of d and k do not allow computing the KL divergence using the kNN method because it requires $k \geq d$.

is a consistent and asymptotically unbiased estimator of the KL divergence as long as $k/D \to 0$, $k \geq d$, and $k \to \infty$ as $D \to \infty$. For large D, the first term is approximately zero. The second and third terms are estimates of the entropy $H(p)$ and the cross-entropy $H_\times(p, q)$.

We first tested this estimator on synthetic data generated from two d dimensional multivariate Gaussian distributions $p = N(-\mu, \mathbf{I})$ and $q = N(\mu, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix and $\mu = \frac{1}{\sqrt{d}} \cdot \mathbf{1}$ with $\mathbf{1}$ being the vector of d ones. Note that $D_{\mathrm{KL}}(p||q) = 2$. Table 1 shows the estimated values from $2 \times 500 - 2 \times 100000$ samples for $d = 1, 10, 100, 200, 300$. The estimates are clearly biased and this bias tends to zero very slowly with increasing number of data samples (it has to go to zero because the estimator (5) is asymptotically unbiased). The bias is due to the estimate of the cross-entropy. While entropy can be estimated accurately even in high-dimensional spaces with small number of data samples, the cross-entropy is harder to estimate. This is because we need to estimate $\log q(x)$ at x where $p(x)$ is still large but we may not have enough data points from $\mathbf{Y}$ in that region. This problem persists for other distributions, such as the Student's t-distribution, which seems to be a relevant model of output from some LSB detectors [23]. With μ approaching zero, the absolute error of the estimate stays approximately the same, producing a very large relative error of the estimated KL divergence. This is quite undesirable because our main interest is to use the benchmark for small payloads when the pdf of covers and stego images are close.

Without any doubts, the KL divergence in the model space is the preferable quantity for benchmarking steganographic schemes because it provides fundamental information about the limits of any steganalysis method. Also, it could be used for evaluating the suitability of models to distinguish between cover and stego objects for a fixed steganographic method (obtaining thus an interesting steganalysis benchmark). It appears, however, that we cannot simply apply existing estimators to data sets that are typically available for steganographic schemes ($d \sim 10 - 300$ and $D \lesssim 10^5$). The effort to remedy this situation could be directed towards deriving better behaved bias-free estimators and reducing the dimensionality of the model space [27].

The problems with the bias of the cross-entropy estimator prompted us to look for alternative statistics for benchmarking that exhibit more stable numerical behavior for sparse data in high-dimensional spaces. We turned our attention to the recently proposed two-sample statistics called Maximum Mean Discrepancy (MMD) [17, 18], which has properties that make it a very good candidate for benchmarking in steganography.

4 Maximum Mean Discrepancy (MMD)

The problem of distinguishing between cover and stego features (3) is a two-sample problem [17]. Assuming the samples $\mathbf{X}$ and $\mathbf{Y}$ were generated from distributions p and q, we need to decide between two hypotheses

$$\begin{aligned} \mathrm{H}_0 &: p = q \\ \mathrm{H}_1 &: p \neq q \ . \end{aligned}$$

From available methods for the two-sample problem (see the review in, e.g., [17]), we decided to use the Maximum Mean Discrepancy (MMD) [17, 18], because of the following advantages relevant to our problem. MMD is numerically stable and scales well with data dimensionality. It has been shown that MMD converges almost independently on data dimension d with error $1/\sqrt{D}$, where D is the number of samples, which allows us to compute an accurate benchmark from $\sim 10^3$ images. Some experimental results on artificial data sets showing this phenomenon will be presented in Section 5 in Table 2. Second, MMD has been well established theoretically and can be linked to other methods, such as Parzen Windows estimates. Third, MMD's computational complexity is $O(D^2)$, which is fast in comparison to Support Vector Machines (SVM), which require expensive grid-search for hyper parameters. We now outline the principles on which MMD is constructed.

To this end, we assume that $\mathcal{X}$ is a separable metric space, and p,q are probability distributions defined on $\mathcal{X}$. The main idea behind MMD is based on Lemma 9.3.2 of [9] stating that $p = q$ if and only if $\forall f \in \mathcal{C}(\mathcal{X})(\mathbf{E}_{\mathsf{x}\sim p} f(\mathsf{x}) = \mathbf{E}_{\mathsf{x}\sim q} f(\mathsf{x}))$, where $\mathcal{C}(\mathcal{X})$ is the class of continuous bounded functions on $\mathcal{X}$. Because this function class is too rich, we cannot use the lemma in finite sample setting. The solution is to restrict the functions to a narrower class $\mathcal{F}$ and measure the disparity between p and q with respect to $\mathcal{F}$ as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f\in\mathcal{F}} \left(\mathbf{E}_{\mathsf{x}\sim p} f(\mathsf{x}) - \mathbf{E}_{\mathsf{y}\sim q} f(\mathsf{y})\right), \tag{6}$$

or in finite sample setting,

$$\mathrm{MMD}[\mathcal{F}, \mathbf{X}, \mathbf{Y}] = \sup_{f\in\mathcal{F}} \left(\frac{1}{D}\sum_{i=1}^{D} f(x_i) - \frac{1}{D}\sum_{i=1}^{D} f(y_i)\right), \tag{7}$$

where $\mathbf{X} = \{x_1, \ldots, x_D\}$, $\mathbf{Y} = \{y_1, \ldots, y_D\}$ are samples (3) from p and q, respectively. To ensure that the measure (6) is useful, we have to choose $\mathcal{F}$

wisely. It has to be rich to distinguish $p \neq q$, yet restrictive enough to provide useful finite sample estimates. The next section shows how to construct such a function class.

4.1 Reproducing Kernel Hilbert Spaces

The class of functions $\mathcal{F}$ used in MMD is built from a symmetric, positive definite[5] function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ called *kernel*. Using the kernel, we define $\forall x \in \mathcal{X}$ the function $K_x : \mathcal{X} \mapsto \mathbb{R}$ as $K_x = k(x, \cdot)$. It is easy to see that the set

$$\mathcal{H}_0 = \{\sum_{i=1}^{n} a_i K_{x_i} | n \in \mathcal{N},\ a_i \in \mathbb{R},\ x_i \in \mathcal{X}\}$$

of all finite linear combinations of K_x, $x \in \mathcal{X}$, forms a vector space of functions $\mathcal{X} \mapsto \mathbb{R}$. The vector space $\mathcal{H}_0$ can be endowed with a dot product defined as

$$\left\langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{j=1}^{m} b_j K_{y_j} \right\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(x_i, y_j). \tag{8}$$

The symmetry and positive definiteness of the kernel function k guarantee that the dot product is well defined and indeed satisfies triangle inequality.

By completing the vector space $\mathcal{H}_0$, we construct Hilbert space $\mathcal{H}$ of real valued functions on $\mathcal{X}$ that can be approximated by finite linear combinations of $k(x, \cdot)$ centered at finite number of points x. This Hilbert space $\mathcal{H}$ has one key property. For each $x \in \mathcal{X}$, the point evaluation functional $\delta_x : \mathcal{H} \mapsto \mathbb{R}$, $\delta_x(f) = f(x)$, is a *continuous* linear functional[6]. This is because $\forall x \in \mathcal{X}$ and $\forall f = \sum a_i K_{x_i} \in \mathcal{H}$

$$\langle f, K_x \rangle_{\mathcal{H}} = \left\langle \sum a_i K_{x_i}, K_x \right\rangle_{\mathcal{H}} = \sum a_i k(x_i, x) = f(x) = \delta_x(f), \tag{9}$$

and the boundedness of δ_x (or continuity) follows from Cauchy-Schwartz inequality $|\delta_x(f)| = |\langle f, K_x \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}}$. Hilbert spaces $\mathcal{H}$ of functions $\mathcal{X} \mapsto \mathbb{R}$ where all point evaluation functionals δ_x are linear and continuous are called *Reproducing Kernel Hilbert Spaces* (RKHS) and (9) is called the reproducing property. Note that since $K_x \in \mathcal{H}$, it can be evaluated it at $y \in \mathcal{X}$ by use of the functional δ_y, which yields

$$\delta_y(K_x) = \langle K_x, K_y \rangle_{\mathcal{H}} = k(x, y). \tag{10}$$

This property makes RKHSs very useful in the theory of SVMs [35].

We can see that an RKHS is tightly linked to its kernel. An important class of kernels are *universal* kernels [40]. We call kernel k universal if $\mathcal{X}$ is compact and its RKHS is dense in $C(\mathcal{X})$ in the maximum (infinity) norm $\|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$. An example of a universal kernel, which we exclusively use in this paper, is the Gaussian kernel on $\mathcal{X} \subset \mathbb{R}^d$

$$k(x, y) = \exp(-\gamma \|x - y\|_2^2),\ \gamma > 0. \tag{11}$$

[5] $k(z_i, z_j)$ is a positive definite matrix for all $l \geq 2$ and all $(z_1, \ldots, z_l), z_i \in \mathcal{X}$.

[6] Convergence in norm in $\mathcal{H}$ implies point-wise convergence.

4.2 MMD

The role of a universal RKHS for MMD will become clear from the following theorem due to [17], which is a simple consequence of Lemma 9.3.2 of [9] and the fact that a universal RKHS is dense in $\mathcal{C}(\mathcal{X})$.

Let $\mathcal{F}$ be a unit ball in a universal RKHS. Then $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$.

The MMD defined over a unit ball in an RKHS accepts a particularly simple form. In a separable Hilbert space, we can exchange expectation and dot product. Thus,

$$\mathbf{E}_{\mathsf{x}\sim p} f(\mathsf{x}) = \mathbf{E}_{\mathsf{x}\sim p} \langle f, K_{\mathsf{x}} \rangle_{\mathcal{H}} = \langle f, \mathbf{E}_{\mathsf{x}\sim p}[K_{\mathsf{x}}] \rangle_{\mathcal{H}} = \langle f, \mu_p \rangle_{\mathcal{H}}\,, \tag{12}$$

assuming the mean value exists $\|\mu_p\|^2_{\mathcal{F}} < \infty$. Thus, the MMD (6) becomes

$$\begin{aligned} \mathrm{MMD}[\mathcal{F}, p, q] &= \sup_{f\in\mathcal{F}} \left(\mathbf{E}_{\mathsf{x}\sim p} f(\mathsf{x}) - \mathbf{E}_{\mathsf{y}\sim q} f(\mathsf{y})\right) = \\ &= \sup_{\|f\|_{\mathcal{F}}\leq 1} \langle f, \mu_p - \mu_q \rangle = \|\mu_p - \mu_q\|_{\mathcal{F}}\,, \end{aligned} \tag{13}$$

because the supremum is reached for $f = (\mu_p - \mu_q)/\|\mu_p - \mu_q\|_{\mathcal{H}}$ from Cauchy-Schwartz inequality. Estimating $\mathrm{MMD}[\mathcal{F}, p, q]$ by replacing μ_p and μ_q in (13) using finite sample estimates $\hat{\mu}_p(x) = \frac{1}{D}\sum_{i=1}^{D} k(x_i, x)$ and $\hat{\mu}_q(x) = \frac{1}{D}\sum_{i=1}^{D} k(y_i, x)$ in (7) leads to a biased estimate. An unbiased estimate based on U-statistics is

$$\mathrm{MMD}_u[\mathcal{F}, \mathbf{X}, \mathbf{Y}] = \left[\frac{1}{D(D-1)}\sum_{i\neq j} k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)\right]^{\frac{1}{2}}. \tag{14}$$

From the theory of U-statistics, under hypothesis H_1 $\mathrm{MMD}^2_u[\mathcal{F}, \mathbf{X}, \mathbf{Y}]$ converges in distribution to a Gaussian according to $\sqrt{D}\left(\mathrm{MMD}^2_u - \mathrm{MMD}^2[\mathcal{F}, p, q]\right) \xrightarrow{\mathcal{D}} N(0, 4\cdot\sigma_u^2)$ [17], where σ_u^2 is the variance of $\mathbf{E}_{\mathsf{x}'\sim p, \mathsf{y}'\sim q}[k(\mathsf{x}, \mathsf{x}') + k(\mathsf{y}, \mathsf{y}') - k(\mathsf{x}, \mathsf{y}') - k(\mathsf{x}', \mathsf{y})]$. The convergence is uniform at rate $1/\sqrt{D}$. The distribution of $\mathrm{MMD}_u[\mathcal{F}, \mathbf{X}, \mathbf{Y}]$ under H_0 can be obtained by bootstrapping (see the recommendations in [18]).

4.3 Analytical calculation of MMD

We now give a specific example of an RKHS generated by the Gaussian kernel (11) $k : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}$, $k(x, y) = \exp\left(-\gamma(x-y)^2\right)$ by providing its orthonormal (ON) basis [41]

$$\left\{ e_n(y) = \sqrt{\frac{(2\gamma)^n}{n!}} y^n \exp(-\gamma y^2) \,\middle|\, n \geq 0 \right\}.$$

Having an ON basis enables us to evaluate the norm in (13) as

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|^2_{\mathcal{F}} = \sum_{n=0}^{\infty} \left(b_{p,n} - b_{q,n}\right)^2,$$

where $b_{p,n} = \langle \mu_p, e_n \rangle_{\mathcal{H}}$ and similarly for q. From (12), (9), and (10)

$$\mu_p(y) = \langle \mu_p, K_y \rangle_{\mathcal{H}} = \mathbf{E}_{\mathsf{x} \sim p} \langle K_{\mathsf{x}}, K_y \rangle_{\mathcal{H}} = \mathbf{E}_{\mathsf{x} \sim p} k(\mathsf{x}, y) = \int_{\mathbb{R}} p(x) \cdot k(x, y) \mathrm{d}x$$
$$= \int_{\mathbb{R}} p(x) \cdot \exp(-\gamma(x-y)^2) \mathrm{d}x = \sum_{n=0}^{\infty} b_{p,n} \sqrt{\frac{(2\gamma)^n}{n!}} y^n \exp(-\gamma y^2).$$

Multiplying the whole equation by $\exp(-\gamma y^2)$, we obtain

$$\int_{\mathbb{R}} p(x) \cdot \exp(-\gamma(x^2 - 2xy)) \mathrm{d}x = \sum_{n=0}^{\infty} b_{p,n} \sqrt{\frac{(2\gamma)^n}{n!}} y^n.$$

From Taylor expansion of function $\int_{\mathbb{R}} p(x) \cdot \exp(-\gamma(x^2 - 2xy)) \mathrm{d}x$ at $y = 0$, we have

$$\sum_{n=0}^{\infty} b_{p,n} \sqrt{\frac{(2\gamma)^n}{n!}} y^n = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n}{\partial y^n} \left[\int_{\mathbb{R}} p(x) \cdot \exp(-\gamma(x^2 - 2xy)) \mathrm{d}x \right] \Big|_{y=0} y^n =$$
$$= \sum_{n=0}^{\infty} \frac{1}{n!} \left[\int_{\mathbb{R}} (2\gamma)^n x^n p(x) \cdot \exp(-\gamma x^2) \mathrm{d}x \right] y^n,$$

and thus

$$b_{p,n} = \int_{\mathbb{R}} p(x) \cdot \sqrt{\frac{(2\gamma)^n}{n!}} x^n \exp(-\gamma x^2) \mathrm{d}x = \int_{\mathbb{R}} p(x) \cdot e_n(x) \mathrm{d}x.$$

The coefficients $b_{p,n}$ are equal to the inner product of p and e_n in L_2.

Extension of this approach to more than one dimension is possible, but quickly becomes computationally intractable. The only exception is when the joint pdf p and q is factorisable $p(x_1, \ldots, x_n) = p(x_1) \cdot \ldots \cdot p(x_d)$ and $q(y_1, \ldots, y_n) = q(x_1) \cdot \ldots \cdot q(x_d)$, in which case it can be easily shown that

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \left(\sum_{n=0}^{\infty} b_{p,n}^2 \right)^d - 2 \left(\sum_{n=0}^{\infty} b_{p,n} b_{q,n} \right)^d + \left(\sum_{n=0}^{\infty} b_{q,n}^2 \right)^d,$$

where $b_{p,n}$, $b_{q,n}$ are as above. This approach was used in Section 5.1 to calculate exact values of MMD for artificially generated data sets.

5 Experiments

In this section, we discuss the choice of the Gaussian kernel parameter γ and then give comparison between the finite sample estimate of MMD (7) with the continuous value (6) on artificial data for various data sample sizes and dimensionality. Finally, we benchmark several popular steganographic techniques with MMD and discuss the results.

Even though universal kernels guarantee that $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$, the choice of the kernel parameter γ has obviously a major influence on the finite sample estimate of MMD (14). If γ is large, the kernel is very narrow and thus $k(x_i, x_j) \approx 0$ (the discrete approximation to the RKHS "overfits" the data). On the other hand, a very small γ leads to a wide kernel and $k(x_i, x_j) \approx 1$ (the approximation is not "pliable" enough). We need the kernel to be aligned with our data[7]. Good results in practice are obtained using the "median" rule [35] (also used in one-class SVMs) according to which γ is set to $\gamma = \frac{1}{\eta^2}$, where η is the median of L_2 divergences between samples. This selection ensures the test statistics to be sensitive to data, because the Gaussian kernel will change its value rapidly.

We also point out that it is important to normalize the data using pre-whitening (setting the data samples to have zero mean and unit variance) before computing the MMD to obtain stable results. Here, we note that any pre-processing we might perform on the data before computing MMD, such as pre-whitening, changes the median and thus the kernel and finally the RKHS. Because for benchmarking of steganography we need one fixed RKHS for all stego methods, we determine the parameters of pre-whitening (and the kernel width γ) from the cover samples only.

5.1 Experiments on artificial data sets

We calculated MMD for the same artificial data from two multinomial Gaussians $N(-\mu, \mathbf{I})$ and $N(\mu, \mathbf{I})$ as in Section 3.4. Table 2 shows that the relative

d	2×500	2×1000	2×5000	2×10000	2×50000	2×100000	MMD
1	−1.29%	−4.17%	−0.57%	1.16%	−0.27%	0.24%	0.562
10	3.79%	2.05%	−0.89%	0.00%	−0.87%	−0.57%	0.123
100	−12.12%	1.71%	−2.41%	−3.50%	0.71%	0.83%	$1.44 \cdot 10^{-2}$
200	1.38%	−4.79%	−1.46%	−2.53%	−0.92%	−0.96%	$7.28 \cdot 10^{-3}$
300	−6.75%	−1.12%	−1.06%	0.48%	0.29%	0.39%	$4.87 \cdot 10^{-3}$

Table 2. Relative error of sample MMD for two d-dimensional multivariate Gaussian distributions computed from D data samples in d-dimensional space.

error of MMD calculated from sample data is remarkably stable across different dimensions d. The sample MMD quickly approaches its theoretical value with increased sample size. Tests with other probability distributions (Laplacian and Student's t-distributions) exhibited very similar convergence rates and errors but are not shown here due to lack of space.

[7] The role of the kernel in MMD is similar to the role of the kernel in Support Vector Machines.

5.2 Benchmarking steganographic methods

In this section, we use MMD to compare statistical detectability of 10 JPEG steganographic algorithms using the 274-dimensional Merged feature set [32]. We focus on low payloads to see if any of the tested steganographic techniques becomes undetectable (indistinguishable using finite sample MMD).

We used a database of 6000 images of a wide variety of scenes from 22 different digital cameras acquired in the raw uncompressed format. The images were embedded with pseudo-random payloads of $5\%, 9\%, 10\%, 15\%$, and 20% bpac (bits per non-zero AC coefficient). The payloads 9% and 10% were chosen intentionally to see the effect of matrix embedding with Hamming codes (the 9% payload can be embedded with a more efficient code). The tested stego algorithms include F5 [42], –F5 [16], F5 without shrinkage [16] (nsF5), JP Hide&Seek[8], Model Based Steganography without deblocking [34] (MBS1), MMx [25], Steghide [19], Perturbed Quantization while double compressing [15](PQ) and its two modifications (PQt and PQe) as described in [16]). The cover images were prepared for every method as if zero message was embedded. The quality factor for the first seven methods was set to 70 and thus the cover images were single-compressed JPEGs with quality 70. Because the three versions of PQ produce double-compressed images, the covers were created by double-compressing the raw images with the same quality factors of 85 and 70.

The empirical estimates $\mathrm{MMD}_u[\mathcal{F}, \mathbf{X}, \mathbf{Y}]$ were calculated from $D = 3000$ examples of $\mathbf{X}$ from the cover class and 3000 examples $\mathbf{Y}$ from the stego class embedded with a specific message length. In each test, the examples were always chosen so that each original raw image appeared either in $\mathbf{X}$ or in $\mathbf{Y}$ but never in both. We always repeated the calculation 100 times with a different split of the 6000 images and took the average as the value of MMD.

To make sure that the MMD was calculated in the same RKHS for every stego method and payload, we determined the whitening parameters and the Gaussian kernel width γ (by the "median" rule) only from the set of cover features $\mathbf{X}$.

In Figure 1 left, we show $-\log_{10}\mathrm{MMD}[\mathcal{F}, \mathbf{X}, \mathbf{Y}]$ for 10 steganographic algorithms and 5 relative payloads. According to this benchmark, the PQ methods and the MMx are the least statistically detectable, while JP Hide&Seek, Steghide, and –F5 are the most detectable. F5 without shrinkage was the best algorithm that does not need side information (the raw image) at the embedder. The horizontal lines mark the value of MMD calculated from two disjoint samples of covers and thus indicate statistical undetectability with respect to the chosen feature set and database. One line is for 70% quality JPEGs for the algorithms producing single-compressed images, while the second line is for double-compressed covers for PQ, PQe, and PQt. We do not show the error bars from the bootstrap because the variances of MMD across different splits of the data set were too small to show in the graph.

To compare the MMD with previously used benchmarks, we show on the right the minimal decision error under equal priors $(P_{FA} + P_{MD})/2$, where P_{FA}

[8] http://linux01.gwdg.de/ alatham/stego.html

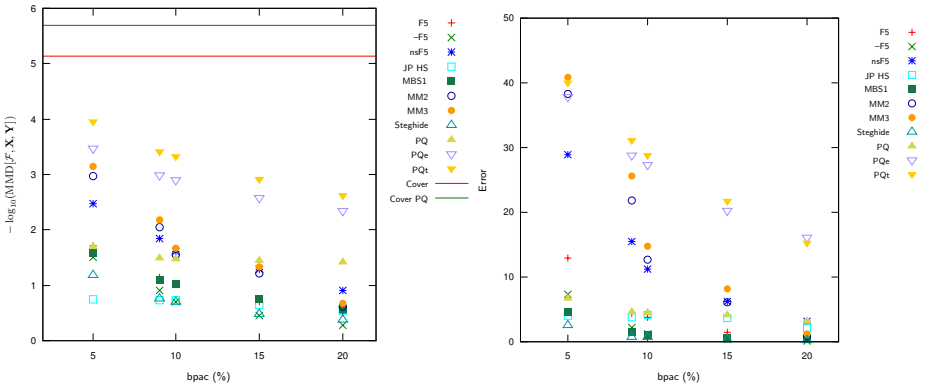

Fig. 1. MMD (left) and probability of error for an SVM (right) for 10 steganographic algorithms and 5 payloads. To obtain a better visual correspondence between the graphs, we show $-\log_{10}\mathrm{MMD}[\mathcal{F},\mathbf{X},\mathbf{Y}]$. The horizontal lines indicate the threshold of undetectability determined as MMD from two samples of covers. Algorithms with MMD close to the line are recognized as secure with respect to the given set of features.

is probability of false positives and P_{MD} the probability of missed detection, for a soft-margin SVM with a Gaussian kernel trained on the same data (one SVM was trained for each payload and method). This quantity was used for benchmarking in [16, 38]. Despite the fact that both benchmarking methods estimate steganographic security in a different way, the graphs appear to be consistent in the sense that stego methods with small MMD tend to have higher classification error and vice versa. We stress that the computational complexity of calculating MMD is significantly smaller than that of training an SVM and calculating the probability of error.

The fast convergence rate and low estimation error even in high-dimensional spaces combined with low computational complexity make the MMD a potentially very useful steganographic benchmark.

6 Conclusions

We proposed a method for benchmarking steganographic schemes. The covers and stego images are first mapped to a feature space, which is viewed as a simplified model of natural images. The statistical detectability of a given method for a fixed payload is then evaluated as a measure of discrepancy between the sample pdf of cover and stego features. As a measure, we investigated the KL divergence and the two-sample statistics called Maximum Mean Discrepancy (MMD). Because the KL divergence is difficult to estimate accurately from sparse data in high-dimensional spaces, we proposed to use the MMD, which has properties useful for applications in steganography. The MMD has a fast convergence rate with respect to the number of data samples even in high-dimensional spaces. Moreover, its computational complexity is proportional to the square of the database

size. The MMD thus replaces the need to train a classifier and enables evaluating statistical detectability from the features themselves. We demonstrate its use on 10 steganographic algorithms and compare the results with a previously used benchmarking method.

The MMD could be also used for benchmarking feature spaces for a fixed steganographic method and thus offers a very interesting approach for comparing steganalytic algorithms. We intend to elaborate on this topic in our future work.

7 Acknowledgments

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under the research grant number FA8750-04-1-0112 and by Air Force Office of Scientific Research under the research grant number FA9550-08-1-0084. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of AFRL, AFOSR, or the U.S. Government.

References

1. I. Avcibas, M. Kharrazi, N. D. Memon, and B. Sankur. Image steganalysis with binary similarity measures. *EURASIP Journal on Applied Signal Processing*, 17:2749–2757, 2005.
2. I. Avcibas, N. D. Memon, and B. Sankur. Steganalysis using image quality metrics. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents III*, volume 4314, pages 523–531, San Jose, CA, January 22–25, 2001.
3. J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen. Non-parametric entropy estimation: An overview. *International Journal of Math. and Stat. Sci.*, 6:17–39, 1997.
4. S. Boltz, E. Debreuve, and M. Barlaud. High-dimensional statistical distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry. In *Proceedings IEEE, Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2007*, pages 1–8, Minneapolis, MN, June 18–23, 2007.
5. C. Cachin. An information-theoretic model for steganography. In D. Aucsmith, editor, *Information Hiding, 2nd International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318, Portland, OR, April 14–17, 1998. Springer-Verlag, New York.
6. R. Chandramouli, M. Kharrazi, and N. D. Memon. Image steganography and steganalysis: Concepts and practice. In T. Kalker, I. J. Cox, and Y. Man Ro, editors, *Digital Watermarking, 2nd International Workshop*, volume 2939 of *Lecture Notes in Computer Science*, pages 35–49, Seoul, Korea, October 20–22, 2003. Springer-Verlag, New York.
7. P. Comesana and F. Pérez-Gonzáles. On the capacity of stegosystems. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.

8. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.
9. R. M. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2002.
10. J. Eggers, R. Bäuml, and B. Girod. A communications approach to steganography. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE Photonic West, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IV*, volume 4675, pages 26–37, San Jose, CA, January 21–24, 2002.
11. H. Farid and L. Siwei. Detecting hidden messages using higher-order statistics and support vector machines. In F. A. P. Petitcolas, editor, *Information Hiding, 5th International Workshop*, volume 2578 of *Lecture Notes in Computer Science*, pages 340–354, Noordwijkerhout, The Netherlands, October 7–9, 2002. Springer-Verlag, New York.
12. E. Franz and A. Schneidewind. Pre-processing for adding noise steganography. In M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, editors, *Information Hiding, 7th International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 189–203, Barcelona, Spain, June 6–8, 2005. Springer-Verlag, Berlin.
13. J. Fridrich. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes. In J. Fridrich, editor, *Information Hiding, 6th International Workshop*, volume 3200 of *Lecture Notes in Computer Science*, pages 67–81, Toronto, Canada, May 23–25, 2004. Springer-Verlag, New York.
14. J. Fridrich, P. Lisoněk, and D. Soukal. On steganographic embedding efficiency. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of *Lecture Notes in Computer Science*, pages 282–296, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.
15. J. Fridrich, M. Goljan, and D. Soukal. Perturbed quantization steganography. *ACM Multimedia System Journal*, 11(2):98–107, 2005.
16. J. Fridrich, T. Pevný, and J. Kodovský. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.
17. A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2007. MPI Technical Report 157.
18. A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.
19. S. Hetzl and P. Mutzel. A graph–theoretic approach to steganography. In J. Dittmann, S. Katzenbeisser, and A. Uhl, editors, *Communications and Multimedia Security, 9th IFIP TC-6 TC-11 International Conference, CMS 2005*, volume 3677 of *Lecture Notes in Computer Science*, pages 119–128, Salzburg, Austria, September 19–21, 2005.
20. A. D. Ker. Steganalysis of LSB matching in grayscale images. *IEEE Signal Processing Letters*, 12(6):441–444, June 2005.
21. A. D. Ker. A capacity result for batch steganography. *IEEE Signal Processing Letters*, 14(8):525–528, 2007.
22. A. D. Ker. The ultimate steganalysis benchmark? In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 141–148, Dallas, TX, September 20–21, 2007.

23. A. D. Ker and R. Böhme. A two-factor error model for quantitative steganalysis. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pages 59–74, San Jose, CA, January 16–19, 2006.
24. M. Kharrazi, H. T. Sencar, and N. D. Memon. Benchmarking steganographic and steganalytic techniques. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 252–263, San Jose, CA, January 16–20, 2005.
25. Y. Kim, Z. Duric, and D. Richards. Modified matrix encoding technique for minimal distortion steganography. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of *Lecture Notes in Computer Science*, pages 314–327, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.
26. S. Lyu and H. Farid. Steganalysis using higher-order image statistics. *IEEE Transactions on Information Forensics and Security*, 1(1):111–119, 2006.
27. Y. Miche, B. Roue, A. Lendasse, and P. Bas. A feature selection methodology for steganalysis. In B. Günsel, A. K. Jain, A. M. Tekalp, and B. Sankur, editors, *Multimedia Content Representation, Classification and Security, InternationalWorkshop*, volume 4105 of *Lecture Notes in Computer Science*, pages 49–56, Istanbul, Turkey, September 11–13, 2006. Springer-Verlag.
28. P. Moulin, M. K. Mihcak, and G. I. Lin. An information–theoretic model for image watermarking and data hiding. In *Proceedings IEEE, International Conference on Image Processing, ICIP 2000*, volume 3, pages 667–670, Vancouver, Canada, September 10–13, 2000.
29. H. Noda, M. Niimi, and E. Kawaguchi. Application of QIM with dead zone for histogram preserving JPEG steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP 2005*, pages II – 1082–5, Genova, Italy, September 11–14, 2005.
30. K. Petrowski, M. Kharrazi, H. T. Sencar, and N. D. Memon. Psteg: Steganographic embedding through patching. In *Proceedings IEEE, International Conference on Acoustics, Speech, and Signal Processing*, pages 537–540, Philadelphia, PA, March 18–23, 2005.
31. T. Pevný and J. Fridrich. Towards multi-class blind steganalyzer for JPEG images. In Mauro Barni, Ingemar J. Cox, Ton Kalker, and Hyoung Joong Kim, editors, *International Workshop on Digital Watermarking*, volume 3710 of *Lecture Notes in Computer Science*, Siena, Italy, September 15–17, 2005. Springer-Verlag, Berlin.
32. T. Pevný and J. Fridrich. Merging Markov and DCT features for multi-class JPEG steganalysis. In E. J. Delp and P. W. Wong, editors, *Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, pages 3 1–3 14, San Jose, CA, January 29 – February 1, 2007.
33. N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, Proceedings of the ACM Symposium on Applied Computing, August 13–17, 2001.
34. P. Sallee. Model-based steganography. In T. Kalker, I. J. Cox, and Y. Man Ro, editors, *Digital Watermarking, 2nd International Workshop*, volume 2939 of *Lecture Notes in Computer Science*, pages 154–167, Seoul, Korea, October 20–22, 2003. Springer-Verlag, New York.
35. B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

36. Y. Q. Shi, C. Chen, and W. Chen. A Markov process based approach to effective attacking JPEG steganography. In J. L. Camenisch, C. S. Collberg, N. F. Johnson, and P. Sallee, editors, *Information Hiding, 8th International Workshop*, volume 4437 of *Lecture Notes in Computer Science*, pages 249–264, Alexandria, VA, July 10–12, 2006. Springer-Verlag, New York.
37. H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *American Journal of Math. and Management Sciences*, 23:301–321, 2003.
38. K. Solanki, A. Sarkar, and B. S. Manjunath. YASS: Yet another steganographic scheme that resists blind steganalysis. In T. Furon, F. Cayre, G. Doërr, and P. Bas, editors, *Information Hiding, 9th International Workshop*, Lecture Notes in Computer Science, pages 16–31, Saint Malo, France, June 11–13, 2007. Springer-Verlag, New York.
39. K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran. Provably secure steganography: Achieving zero K-L divergence using statistical restoration. In *Proceedings IEEE, International Conference on Image Processing, ICIP 2006*, pages 125–128, Atlanta, GA, October 8–11, 2006.
40. I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.
41. I. Steinwart, D. Hush, and C. Scovel. An explicit description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52:4635–4643, 2006. Los Alamos National Laboratory Technical Report LA-UR-04-8274.
42. A. Westfeld. High capacity despite better steganalysis (F5 – a steganographic algorithm). In I. S. Moskowitz, editor, *Information Hiding, 4th International Workshop*, volume 2137 of *Lecture Notes in Computer Science*, pages 289–302, Pittsburgh, PA, April 25–27, 2001. Springer-Verlag, New York.
43. G. Xuan, Y. Q. Shi, J. Gao, D. Zou, C. Yang, Z. Z. P. Chai, C. Chen, and W. Chen. Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions. In M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, editors, *Information Hiding, 7th International Workshop*, volume 3727 of *Lecture Notes in Computer Science*, pages 262–277, Barcelona, Spain, June 6–8, 2005. Springer-Verlag, Berlin.