Applications of Explicit Non-Linear Feature Maps in Steganalysis

Mehdi Boroumand, *Student Member, IEEE*, and Jessica Fridrich, *Fellow, IEEE*

***Abstract*—Currently, the most popular detectors of content-adaptive image steganography are built using machine learning with images represented with rich features. Such high-dimensional descriptors, however, prevent utilization of more complex and potentially more accurate machine learning paradigms, such as kernelized support vector machines, due to infeasibly expensive training. In this paper, we demonstrate that explicit non-linear feature maps coupled with simple classifiers improve the accuracy of current steganalysis detectors built as binary classifiers as well as quantitative detectors in the form of payload regressors. The non-linear map is obtained by approximating a symmetric positive semi-definite kernel on selected pairs of cover features. Exponential forms of kernels derived from symmetrized Ali-Silvey distances improve the detection accuracy of binary detectors and lower the error of quantitative detectors across all tested steganographic schemes on grayscale and color images. The learned non-linear map only weakly depends on the cover source and its learning has a low computational complexity. The technique can also be used for unsupervised feature dimensionality reduction. For payload regressors, the dimensionality can be significantly reduced while simultaneously decreasing the estimation error.**

***Index Terms*—Steganalysis, adaptive steganography, explicit transformation, Nyström approximation, support vector machine**

I. Introduction

The prevailing paradigm for detection of content-adaptive steganography [1]–[7] uses machine learning with high dimensional image representations called rich media models [8]–[21].[1] This holds true for binary detectors targeted to a specific steganographic algorithm as well as quantitative detectors that estimate the number of embedding changes (payload size) [24], [25]. The high dimensionality of rich models, however, limits the choices for the detector construction due to prohibitively expensive training. The problem becomes worse for quantitative detectors estimating the payload size [24] as support vector regression (SVR) generally requires an expensive three-dimensional search for the hyperparameters. The community has thus sacrificed some of the detection accuracy by resorting to simpler machine learning tools, such as the ensemble of FLD base learners [26] (Fisher Linear Discriminant), its linear version derived as a likelihood ratio test based on a multivariate Gaussian model of base learners' projections [27], regularized linear discriminant [28] implemented using the large linear system optimization method called LSMR [29], regression trees [25] used for estimating payload size, and the on-line average ensemble perceptron [30] for large training sets.

Prior art offers evidence that the accuracy of steganalysis can be improved whenever it is feasible to train the detector as a Gaussian support vector machine (G-SVM).[2] In [8], the authors pointed out that a G-SVM trained on a 3,300-dimensional subset of the 12,753-dimensional spatial rich model (SRMQ1) selected using a simple forward feature selection method on SRM submodels performed better than the FLD-ensemble with the rich model (see Table II in [8]). Feature sets of dimensionality less than 2,000 formed by variable quantization co-occurrences with a G-SVM were also shown to be competitive with detectors built using rich models and the FLD-ensemble [11].

The approach proposed in this paper finds a non-linear decision boundary between cover and stego features by training a simple (e.g., linear) classifier on features that were transformed using an explicit non-linear mapping. The mapping is obtained by approximating an implicit feature transformation determined by a symmetric positive semi-definite kernel and can be interpreted as endowing the Euclidean feature space with a different metric. This well-founded approach has been used in machine learning and specifically in computer vision as a low-cost method for training kernelized SVMs [33], [34]. Square-rooting features, which has been employed for object retrieval [35] and in digital forensics [36], is a special case of the methodology introduced in this paper.

The value of the proposed approach is demonstrated on four content-adaptive embedding algorithms in the spatial domain on standard image sets with grayscale as well as

The work on this paper was supported by the Air Force Office of Scientific Research under the research grant FA9550-09-1-0147. 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 AFOSR or the U.S. Government.

The following authors are with the Department of Electrical and Computer Engineering, Binghamton University, NY, 13902, USA. Email: {mboroum1,fridrich}@binghamton.edu

[1]Recently, convolutional neural networks [22], [23] have been proposed as an alternative to this established detection paradigm.

[2]In the JPEG domain, steganalysis with co-occurrences of quantized DCT coefficients [31] does not seem to benefit from kernelized SVMs because the classes of cover and stego features are approximately linearly separable [32], mostly likely because such features are built directly in the embedding domain.

color images for both binary and quantitative steganography detectors. The non-linear mapping can also be used for unsupervised reduction of feature dimensionality. For quantitative detectors, the reduced features surprisingly further decrease the payload estimation error.

In the next section, we provide background information regarding kernels derived from symmetrized Ali–Silvey distances, the associated explicit feature transformation, and its Nyström approximation learned from a set of training feature vectors. We also explain how to extend the methodology to high-dimensional rich models with a low computational complexity and memory requirements. After spelling out the common core of experiments in Section III, we experimentally investigate the suitability of various kernels for steganalysis on low-dimensional feature sets in Section IV. The results of all experiments on color and grayscale images with binary detectors appear in Section VI, while Section VII is devoted to experiments with quantitative detectors estimating the payload size. Since the proposed feature transformation can be used for unsupervised dimensionality reduction, in Section VIII we report how the detection accuracy of binary classifiers and the estimation error of payload size regressors depends on the number of retained dimensions of the transformed features. After discussing the limitations of the proposed method in Section IX, the paper is concluded in Section X.

This paper is an extension of our initial investigation [37]. In contrast to this workshop version, here we frame the methodology more rigorously within the body of existing literature on approximations of explicit feature maps. Second, we extended the approach from binary classifiers to quantitative detectors (payload size regressors) constructed using regression trees as well as linear regressors. Third, we expanded the experiments with binary classifiers to color images on the never-before-studied version of color features that consider the selection channel (the maxSCRM). Fourth, we substantially expanded the section dealing with dimensionality reduction and its effect on detection accuracy of binary classifiers as well as payload regressors. Surprisingly, reducing the feature dimensionality by 60% further decreases the payload estimation error of quantitative detectors.

For better readability, below we provide a list of frequently used symbols.

M	Number of images for map training
D	Feature dimensionality
N_{trn}	Number of images for detector training
$\mathbf{x}^{(i)}$, $\mathbf{y}^{(i)}$	Cover, stego features
$\varphi, \varphi^{(\mathrm{h})}$	Explicit map, its component form
E	Number of retained dimensions in range of φ
$k, k^{(\mathrm{h})}$	Kernel, its component form
$d(\mathbf{x}, \mathbf{y})$	Distance between vectors $\mathbf{x}, \mathbf{y}$
$\langle \cdot, \cdot \rangle$	Dot product
P_{E}	Minimal total detection error

II. Transformations from kernels

Virtually all rich models of images used for steganalysis in the spatial as well as JPEG domain are either histograms or higher-order co-occurrences of noise residuals or DCT coefficients or their differences. A D-dimensional feature vector $\mathbf{x}$ extracted from an image is thus an element of $\mathbb{R}_+^D$, where $\mathbb{R}_+$ stands for the set of non-negative real numbers.

A. Kernels

Mapping $k : \mathbb{R}_+^D \times \mathbb{R}_+^D \to \mathbb{R}_+$ is called positive semi-definite if for any n and any $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)} \in \mathbb{R}_+^D$, the $n \times n$ matrix $K_{ij} = k(\mathbf{u}^{(i)}, \mathbf{u}^{(j)})$ is positive semi-definite. Many symmetric positive semi-definite kernels used in computer vision are either additive or multiplicative: $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D k^{(\mathrm{h})}(x_i, y_i)$ and $k(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^D k^{(\mathrm{h})}(x_i, y_i)$, respectively. The component kernel $k^{(\mathrm{h})} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ is called γ-homogeneous if $k^{(\mathrm{h})}(cx, cy) = c^\gamma k^{(\mathrm{h})}(x, y)$ for all $c \geq 0$ and stationary if $k^{(\mathrm{h})}(c+x, c+y) = k^{(\mathrm{h})}(x, y)$ for all c. We now introduce several kernels commonly used in machine learning that will be studied in this paper in the context of steganalysis.

A kernel k defines a metric (distance) in $\mathbb{R}_+^D$ using the formula

$$d^2(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{y}, \mathbf{y}) - 2k(\mathbf{x}, \mathbf{y}). \tag{1}$$

For example, assuming that vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ are L_2-normalized, meaning that $\|\mathbf{x}\|_2^2 = \sum_{i=1}^D x_i^2 = \|\mathbf{y}\|_2^2 = 1$, their square Euclidean distance can be written as $\|\mathbf{x} - \mathbf{y}\|_2^2 = 2\,(1 - k(\mathbf{x}, \mathbf{y}))$, where we introduced

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D x_i y_i \tag{2}$$

for the dot product, a homogeneous additive **linear kernel** with $k^{(\mathrm{h})}(x, y) = xy$. In this case, and in general, the value $k(\mathbf{x}, \mathbf{y})$ can be thought of as a measure of similarity. An entire family of kernels can be constructed using a similar approach from symmetrized Ali–Silvey distance measures [38] also called f-divergences that share the leading term with the Kullback–Leibler divergence in the limit when the class distributions are close.

The **Hellinger kernel**

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D \sqrt{x_i y_i}, \tag{3}$$

is derived from the Bhattacharyya distance. Here, $\mathbf{x}$ and $\mathbf{y}$ need to be L_1-normalized. Note that the Hellinger kernel corresponds to the linear kernel on square-rooted features.

The **chi-square kernel**

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D \frac{2x_i y_i}{x_i + y_i} \tag{4}$$

with $\mathbf{x}$ and $\mathbf{y}$ L_1-normalized originates from the χ^2 distance defined as the symmetrized χ^2 statistic:

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= \frac{1}{4} \sum_{i=1}^D \frac{(x_i - y_i)^2}{(x_i + y_i)/2} = \frac{1}{2} \sum_{i=1}^D \frac{(x_i + y_i)^2 - 4x_i y_i}{x_i + y_i} \\ &= 1 - \sum_{i=1}^D \frac{2x_i y_i}{x_i + y_i} = 1 - k(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{5}$$

with $\mathbf{x}$ and $\mathbf{y}$ L_1-normalized.

The **Jensen–Shannon kernel**

$$k(\mathbf{x},\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{D} x_i \log\frac{x_i+y_i}{x_i} + y_i \log\frac{x_i+y_i}{y_i} \tag{6}$$

is easily shown to stem from the symmetrized KL divergence

$$d^2(\mathbf{x},\mathbf{y}) = D_{\mathrm{KL}}\left(\mathbf{x}\middle\|\frac{\mathbf{x}+\mathbf{y}}{2}\right) + D_{\mathrm{KL}}\left(\mathbf{y}\middle\|\frac{\mathbf{x}+\mathbf{y}}{2}\right) \tag{7}$$

when $\mathbf{x}$ and $\mathbf{y}$ are L_1-normalized.

Since for a symmetric positive semi-definite kernel k and $\alpha > 0$, $e^{\frac{1}{\alpha}(k(\mathbf{x},\mathbf{y})-1)}$, is also symmetric positive semi-definite, it can be used to define exponential counterparts of all the above kernels. We will refer to them using the preposition 'exp', such as 'exp-linear.' Note that the exponential form of an additive kernel becomes multiplicative. Also note that, due to normalization, $0 \le k(\mathbf{x},\mathbf{y}) \le 1$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}_+^D$ and all kernels. Undefined terms in the chi-square and the Jensen–Shannon kernels due to division by zero or logarithm of zero are set to zero.

B. Explicit feature embedding

A kernelized SVM is a linear SVM in a space of features transformed into a Hilbert space $\mathcal{H}$, $\varphi : \mathbb{R}_+^D \to \mathcal{H}$, endowed with a dot product $\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H} \to \mathbb{R}_+$ such that

$$\langle\varphi(\mathbf{x}),\varphi(\mathbf{y})\rangle = k(\mathbf{x},\mathbf{y}) \text{ for all } \mathbf{x},\mathbf{y} \in \mathbb{R}_+^D. \tag{8}$$

In the classic exposition to SVMs [39], the transformation φ is implicit and does not need to be constructed because the SVM can be trained and evaluated by using dot products in the Hilbert space, which can be computed using the kernel (8). This has become recognized as the "kernel trick." However, despite the availability of efficient solvers for linear SVMs, with the training set size N_{trn} the complexity of training a kernelized SVM is $\mathcal{O}(\max\{N_{trn},D\}\times\min\{N_{trn},D\}^2)$ whether the primal or dual formulation is used [40]. This is a problem in steganalysis since the feature dimensionality of rich models is $D \approx 10^4$ and the training sets also need to be large (10^4 or larger).

One possibility to decrease the training complexity is to use an explicit form of φ truncated to a finite-dimensional space, $\hat{\varphi} : \mathbb{R}_+^D \to \mathbb{R}^E$, where one can make use of classifiers with low training complexity, such as the FLD-ensemble, on transformed features $\hat{\varphi}(\mathbf{x}^{(i)})$ and $\hat{\varphi}(\mathbf{y}^{(i)})$, $i = 1,\ldots,N_{trn}$. To classify a feature $\mathbf{z} \in \mathbb{R}_+^D$, the low-complexity classifier is presented with $\hat{\varphi}(\mathbf{z})$. Next, we review several possibilities for finding an approximation $\hat{\varphi}$ to the explicit form of φ and identify the one that will be used in this paper.

The map that satisfies (8) can be found explicitly using Bochner's Theorem [33], [41]. For component kernel $k^{(\mathrm{h})}(x,y)$, the component map $\varphi^{(\mathrm{h})}$ assigns a complex-valued function Ψ (a function of ω) to each component x

$$\Psi_\omega(x) = \exp(-i\omega\log x)\sqrt{x^\gamma\kappa(\omega)} \tag{9}$$

$$\Psi_\omega(x) = \exp(-i\omega)\sqrt{\kappa(\omega)} \tag{10}$$

with $\omega \in \mathbb{R}$ and $\kappa(\omega)$ being the kernel spectrum, which adopts a rather simple closed form for most commonly used kernels (see, e.g., Fig. 1 in [33]). The form of the mapping in (9) and (10) corresponds to γ-homogeneous and stationary kernels, respectively. We note that $k^{(\mathrm{h})}(x,y) = \int_{\mathbb{R}} \Psi_\omega^\star(x)\Psi_\omega(x)\mathrm{d}\omega = \left\langle\varphi^{(\mathrm{h})}(x),\varphi^{(\mathrm{h})}(y)\right\rangle$, where $\Psi^\star$ denotes the complex conjugate. The mapping for the kernel k operating on D-dimensional vectors is obtained for additive and multiplicative kernels by

$$\varphi(\mathbf{x}) = \bigoplus_{i=1}^{D}\varphi^{(\mathrm{h})}(x_i), \quad \varphi(\mathbf{x}) = \bigotimes_{i=1}^{D}\varphi^{(\mathrm{h})}(x_i), \tag{11}$$

where the operators $\bigoplus$ and $\bigotimes$ stand for cartesian and Kronecker product. In particular, if the range of $\varphi^{(\mathrm{h})}$ is constrained to $\mathbb{R}^E$, the dimension of $\varphi(\mathbf{x})$ is $E\times D$ and E^D, respectively.

Since φ maps to an infinite dimensional space of complex-valued functions, discretization (approximation) of some form, $\hat{\varphi}$, is needed for practical applications. This can be achieved by either requesting that the range of features be a compact set (which makes the kernel spectrum discrete) or by imposing periodicity on the kernel [33]. The complexity of this approach associated with multiplicative kernels (Section 7 in [33]), however, lead us to the third option, which is the Nyström approximation. It has the advantage of being computationally rather efficient for both additive and multiplicative kernels and it is simple to implement. On the other hand, it requires a training set, which makes the mapping $\hat{\varphi}$ data dependent. According to our experience, this dependence is weak, and the training set can be rather small even for high-dimensional rich features.

C. Nyström approximation

We opted for an accessible explanation of this method adapted from [34]. The approximation of φ, which we denote with $\hat{\varphi} : \mathbb{R}_+^D \to \mathbb{R}^M$, will be derived from M ($D \le M \le N_{trn}$) training vectors $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)} \in \mathbb{R}_+^D$. Specifically, we find the images of $\mathbf{x}^{(i)}$ under $\hat{\varphi}$, $\hat{\varphi}(\mathbf{x}^{(i)})$, that approximate the kernel evaluated on all pairs of training vectors by minimizing

$$\sum_{i,j=1}^{M}\left(k(\mathbf{x}^{(i)},\mathbf{x}^{(j)}) - \hat{\varphi}(\mathbf{x}^{(i)})\cdot\hat{\varphi}(\mathbf{x}^{(j)})\right)^2. \tag{12}$$

To avoid redundant components, the following orthogonality constraint is added to the minimization problem:

$$\sum_{i=1}^{M}\hat{\varphi}_a(\mathbf{x}^{(i)})\hat{\varphi}_b(\mathbf{x}^{(i)}) = 0 \text{ for all } 0 \le a \ne b \le M, \tag{13}$$

where $\hat{\varphi}_a(\mathbf{x})$, $1 \le a \le M$ denotes the ath component of $\hat{\varphi}(\mathbf{x}) \in \mathbb{R}^M$. This constrained optimization can be solved using the method of Lagrange multipliers, which gives that the ath coordinate of vector $\hat{\varphi}(\mathbf{x}^{(i)})$ across i, $\phi_a \triangleq (\phi_a(\mathbf{x}^{(1)}), \dots, \phi_a(\mathbf{x}^{(M)}))' \in \mathbb{R}^M$, must be an eigenvector of the kernel matrix $\mathbf{K} = (K_{ij}) \in \mathbb{R}_+^{M\times M}$, $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$:

$$\mathbf{K}\phi_a = \lambda_a^2 \phi_a,\ 1 \le a \le M, \tag{14}$$

with the components permuted if necessary to have the eigenvalues λ_a^2 non-increasing: $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_M$. We note that $\lambda_a = \|\phi_a\|_2$.

Having determined the images of training vectors $\mathbf{x}^{(i)}$, $\hat{\varphi}(\mathbf{x}^{(i)})$, the Nyström approximation of $\varphi(\mathbf{z})$ for any $\mathbf{z} \in \mathbb{R}_+^D$ is a linear combination of transformed vectors $\hat{\varphi}(\mathbf{x}^{(i)})$ with coefficients given by the "projections" of $\mathbf{z}$ on $\mathbf{x}^{(i)}$ evaluated using the kernel $\mathbf{K}(\mathbf{z}, \cdot) = \big(k(\mathbf{z}, \mathbf{x}^{(1)}), \dots, k(\mathbf{z}, \mathbf{x}^{(M)})\big)$. Formally, $\hat{\varphi}(\mathbf{z}) = (\hat{\varphi}_1(\mathbf{z}), \dots, \hat{\varphi}_M(\mathbf{z}))$, with

$$\hat{\varphi}_a(\mathbf{z}) = \frac{1}{\lambda_a^2}\mathbf{K}(\mathbf{z}, \cdot) \cdot \phi_a,\ 1 \le a \le M. \tag{15}$$

By retaining the first $E \le M$ coordinates a corresponding to the largest eigenvalues λ_a^2, we can restrict the dimensionality of the explicit map $\hat{\varphi} : \mathbb{R}_+^D \to \mathbb{R}^E$. When $E = D$, the feature transform $\hat{\varphi}$ preserves the feature dimensionality.

Because the Hellinger kernel (3) corresponds to the linear kernel (2) on L_1-normalized features that have been square-rooted elementwise, the explicit map $\hat{\varphi}$ for this kernel adopts a particularly simple closed form:

$$\hat{\varphi}(\mathbf{z}) = (\sqrt{z_1}, \dots, \sqrt{z_E}). \tag{16}$$

The complexity of this particular feature transform (16) is negligible in comparison to the classifier training. This is why in our experiments, we include results obtained with square-rooted features. They are expected to match the results obtained with the transform learned using the above Nyström approximation for the Hellinger kernel.

III. Common core of experiments

Unless stated otherwise, all experiments in this paper were conducted with four steganographic methods, WOW [2], S-UNIWARD [4], HILL [5], and MiPOD [7] on BOSSbase 1.01 [42] containing 10,000 512×512 8-bit grayscale images and on its color version, called BOSSColor, with 24-bit PPM images prepared using the same script that was used for creating BOSSbase with the following modifications [20]. The output format of 'ufraw' (ver. 0.18 with 'dcraw' ver. 9.06) was changed to PPM and all calls of 'convert' used PPM for the output as well as for Lanczos resizing so that the smaller image dimension was 512 before central cropping to 512×512.

Experiments were run on ten random splits of the database into 5000 training and 5000 testing images. The detection performance was evaluated with the minimum total error probability $P_{\text{E}} = \min_{P_{\text{FA}}} \frac{1}{2}(P_{\text{FA}} + P_{\text{MD}})$, where P_{FA} and P_{MD} are the empirical false-alarm and missed-detection rates, averaged over the ten splits, $\overline{P}_{\text{E}}$. The performance of payload regressors was evaluated using mean square error (MSE) and mean absolute error (MAE) on the testing set, again averaged over ten splits.

The feature transformation was learned on M randomly selected cover images from the training set. Unless mentioned otherwise, the FLD-ensemble (or the regressor) was then trained and tested on the transformed features. The constant α in exponential kernels was computed from M training features $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}$ as $\alpha = 1/M^2 \sum_{i,j=1}^{M} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

Experiments on grayscale images were purposely conducted with maxSRMd2 features [13] to obtain the best possible performance within the paradigm of steganalysis with rich models. For color images, we used the spatio-color rich model (SCRM) [20], which is a merger of the SRMQ1 (SRM with residual quantization $q = 1$) and the color rich model (CRM). We modified the feature set to maxSCRM that incorporates the knowledge of the selection channel as in [13]. Note that, according to the best knowledge of the authors, this is the first time a selection-channel-aware feature set was used for detection of steganography in color images. Again, we opted for this more powerful version of the SCRM in order to report the best possible detection with rich models.

In all our experiments, we removed from all rich models those co-occurrence bins that are equal to zero independently of the input image. We did so to stabilize the numerical computations and prevent ill-conditioned matrices in the eigenvector problem (14). The empty bins, which are due to the overlap of adjacent residual supports, occur only in the 'minmax22h', 'minmax34h', and 'minmax41' submodels of SRM, maxSRM, and SRMQ1 for first-order and third-order differences. The CRM part of the color SCRM does not contain any zero features because the supports of color residuals are disjoint as they are computed from different color channels. For SRM and SRMQ1, only 101 bins in the above-mentioned submodels are generally non zero, which makes the true dimensionality of SRMQ1 equal to $11{,}409 = 12{,}753 - 2\times3\times(325-101)$ and the dimensionality of SCRMQ1 equal to $16{,}813 = 11{,}408 + 5{,}404$.

Because maxSRMd2 [13] uses the 'd2' scan for forming co-occurrences (residuals with indices (i, j), $(i, j+1)$, $(i+1, j+2)$, $(i+1, j+3)$ and three more horizontally and vertically flipped versions), the number of non-zero elements in the above three submodels is different. For quantization $q = 2$, the 'minmax22h', 'minmax34h', and 'minmax41' submodels for the first and third order residuals have dimensionality 190. For $q = 1$ and $q = 1.5$, their dimensionality is 120. This gives the maxSRMd2 feature set a dimensionality of 32,016.

IV. Pilot study

The purpose of this section is to study the benefits of using the non-linear transformation on individual submodels of the SRMQ1 rich model and to assess the effectiveness of different kernels introduced in Section II-A.

The first experiment was carried out with S-UNIWARD [4] and HILL [5] with a fixed payload of 0.4 bits per pixel (bpp) and two low dimensional feature sets – the 169-dimensional co-occurrence matrix obtained using the "KB residual" from the 'SQUARE 3x3' submodel and the 'minmax22h' submodel for the first-order residual, both quantized with quantization step $q = 1$. The 'minmax22h' submodel has effective dimensionality of only 101 after removing from it elements that are always equal to zero. Figure 1 shows the detection error $\overline{P}_{\rm E}$ when the FLD-ensemble was trained on the original form of the features, their square rooted form, and features transformed with the Nyström approximation of eight kernels. For comparison, we also report the results with a G-SVM trained on the original features. The exponential versions of kernels are consistently better than their additive counterparts and offer similar detection gain w.r.t. non-transformed features (up to 3.5%). While the G-SVM always offers the best performance, the exponential kernels match it except for the 'minmax22h' feature set. Finally, we confirm that the results obtained with the Nyström approximation of the Hellinger kernel indeed match the square-rooted features. The statistical spread of the results in terms of the mean absolute deviation was in the range $0.0016 - 0.0039$.

In the next experiment, we assessed the effectiveness of the non-linear mapping across *all* submodels of the SRMQ1. This was executed for the embedding algorithm WOW at 0.4 bpp. Because the detection accuracy of individual submodels varies substantially, in Figure 2 we show the difference between $P_{\rm E}$ obtained using the non-transformed features / transformed features and $P_{\rm E}$ obtained using a Gaussian SVM for that submodel. The transformation decreases the detection error by up to 4% (s1-minmax34v), depending on the submodel. For some submodels, the transformed features can perform equally well as the G-SVM, e.g., for s35-spam11 and s5 × 5-minmax22v. Certain submodels do not benefit from the transform at all: s1-minmax41, s1-minmax34h, and s3-minmax34. The improvement in detection error will necessarily depend on the degree of non-linearity of the decision boundary between cover and stego features and on the accuracy of the learned mapping $\hat{\varphi}$ as an approximation of the infinite-dimensional transformation defined in (9)–(10).

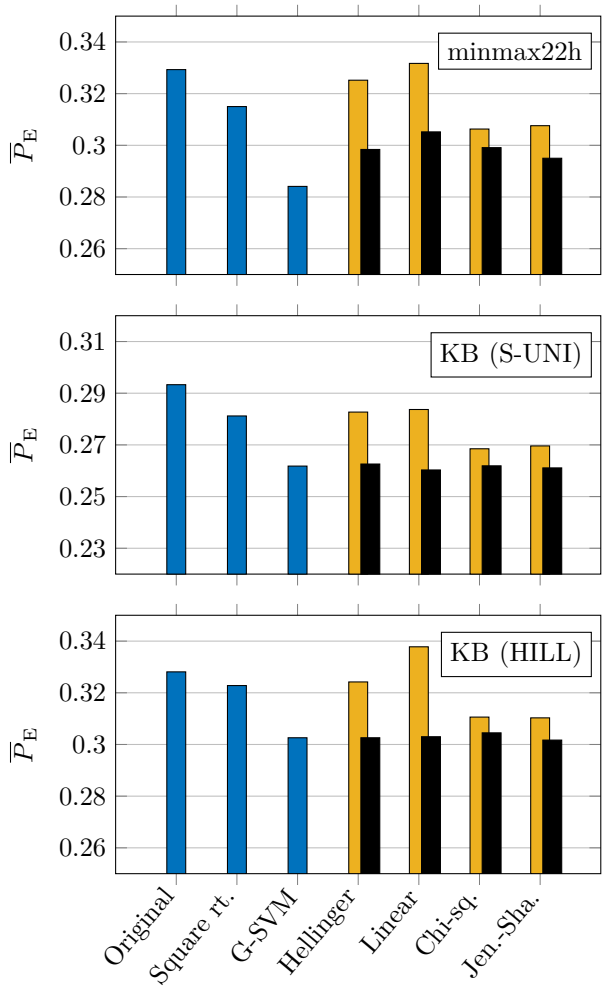

Figure 1. $\overline{P}_{\rm E}$ for various kernels for S-UNIWARD and HILL at 0.4 bpp for KB co-occurrence and 'minmax22h'. The exponential versions of kernels are the overlapping black bars.

V. Transforming rich models

The complexity of learning the map is largely determined by the training set size, M, because forming the $M \times M$ matrix $\mathbf{K}$ requires $\mathcal{O}(DM^2)$ operations and the complexity of the eigenvector problem (14) is $\mathcal{O}(M^3)$ if implemented using the Cholesky decomposition.[3] This makes the total training complexity $\mathcal{O}(DM^2 + M^3)$. The eigenvector problem (14) only needs to be executed once for a given cover source because the map is trained on cover images only and is thus independent of the payload size and the steganographic scheme. Once the eigenvectors are found, the cost of transforming a new feature $\mathbf{z} \in \mathbb{R}_+^D$ is $\mathcal{O}(MD)$ to evaluate $\mathbf{K}(\mathbf{z}, \cdot)$ and $\mathcal{O}(ME)$ to compute all E coordinates of $\varphi(\mathbf{z})$ (15).

Because we need $M \geq D \approx 10^4$ training images for typical rich feature sets, the complexity of training the non-linear map for the entire feature vector, $\mathcal{O}(DM^2 + M^3)$, would be infeasibly large. Moreover, the constraint $M \geq D$ implies that we would not be able to benchmark the performance on standard image sets, such as BOSSbase.

Thus, instead of applying the Nyström approximation directly to rich models, we learned the map for each submodel of the rich model. For this task, the same M training images are used across all submodels. Everywhere in this paper, the mapping $\hat{\varphi}$ was trained only on $M = 350$ cover images as we did not observe any gain when increasing M. In fact, based on [43], [44] we hypothesize that it may be possible to train the map on a set of "canonical" features that represent a large variety of images or even on synthetic features.

[3]Note that Matlab's implementation of `'eig.m'` is based on an iterative algorithm with a different complexity.

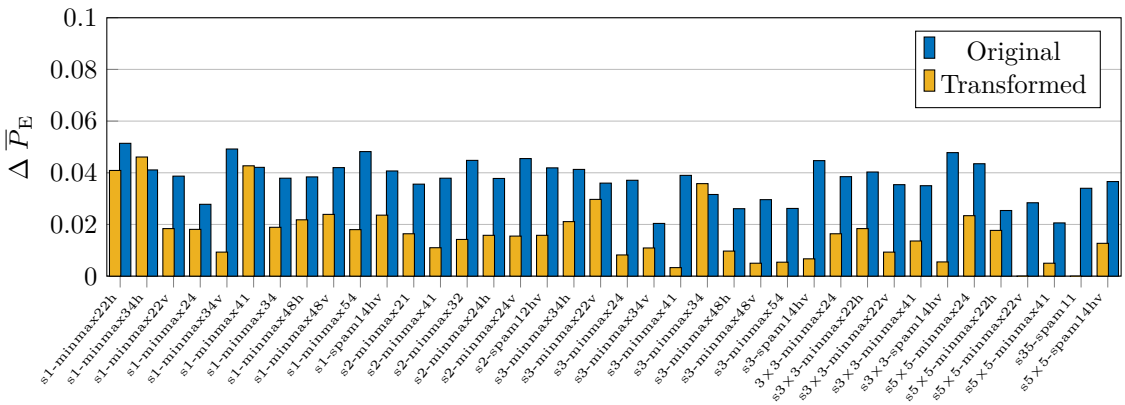

Figure 2. Difference in P_{E} obtained with the non-transformed / transformed features and P_{E} obtained with a G-SVM for all submodels of SRMQ1. WOW at 0.4 bpp.

VI. Binary detectors

In this section, we report the results of all experiments with steganography detectors implemented as binary classifiers on rich models. In order not to increase the feature dimensionality, the number of retained coordinates was kept at $E = d$, where d is the submodel dimensionality.

For each payload, a separate FLD-ensemble [26] was trained on the original features and on the transformed features. For each split of the database into a training and testing set, the map $\hat{\varphi}$ was retrained. Only the exponential Hellinger was tested because it provided a similar performance as the other exponential kernels in Section IV and because it has lower complexity than the chi-square or Jensen–Shannon kernels.

A. Grayscale images

Table I shows $\overline{P}_{\mathrm{E}}$ as a function of payload for the original maxSRMd2 feature set [13], its square rooted version, and the transformed version using exp-Hellinger on BOSSbase 1.01. A consistent improvement of up to $\approx 3\%$ in $\overline{P}_{\mathrm{E}}$ is obtained across all four embedding algorithms with the biggest gain for WOW and the smallest for MiPOD and HILL. A qualitatively similar behavior and gains were observed with the SRM features while all detection errors were merely slightly larger, which is why we do not report these essentially redundant results in this paper. We have also investigated the gain by replacing the FLD-ensemble with the regularized linear discriminant implemented using the LSMR optimizer [28]. The results were in majority of cases within the statistical spread of the results obtained with the ensemble and are thus not included in the paper.

B. Color images

In our second experiment, the above four steganographic algorithms were applied by color channels to images from BOSSColor (Section III) by embedding the same relative payload in each channel. The complete results are listed in Table II. The non-linear map boosts detection to a different degree depending on the steganographic method and payload. The largest gain of almost 4% is observed for WOW for medium payloads.

VII. Quantitative steganalysis

A general approach to quantitative steganalysis proposed in [24] calls for a regressor between a feature extracted from an image and the payload. The same publication demonstrated the benefit of using non-linear regressors (SVRs) over linear ones when detecting LSB matching. With modern content-adaptive schemes, the dependence of the feature on payload is likely going to be much more complex (and non-linear) than for non-adaptive LSB matching since the shift of the feature vector due to embedding is likely more influenced by content. The complexity of training a SVR with high-dimensional features forced researchers to look into alternative non-linear regressors whose complexity better scales with feature dimensionality, such as regression trees [25]. In this section, we report that both linear regressors and regression trees benefit from explicit non-linear feature transforms.

Experiments were conducted again with all four steganographic schemes on BOSSbase images. To lower the computational complexity of all experiments, we used the SRMQ1 feature vector. The stego images used for training the regressor were embedded with relative payload size R selected uniformly randomly from $[0, 1]$ bpp. Table III lists the MSE and MAE for linear regressors and regression trees for all four tested steganographic schemes. Figure 3 contains a few examples of scatter plots showing the estimated payload $\hat{R}$ vs. the true payload R when training the regressors on the original features and the transformed features. Overall, the non-linear feature transformation decreases the MSE of linear regressors by 13–22% and regression trees by 4–14%. The biggest gain was observed for S-UNIWARD and the smallest for HILL and MiPOD. The scatter plots show how the non-linear map redistributes the payload estimation error.

Table I
DETECTION ERROR $\overline{P}_{\mathrm{E}}$ FOR FOUR STEGANOGRAPHIC SCHEMES AND A RANGE OF PAYLOADS ON BOSSBASE 1.01. THE CLASSIFIER WAS THE FLD-ENSEMBLE TRAINED WITH MAXSRMD2 FEATURES, THEIR SQUARE ROOTED FORM, AND TRANSFORMED BY SUBMODELS USING NYSTRÖM APPROXIMATION OF THE EXPONENTIAL HELLINGER KERNEL.

	Payload (bits per pixel)					
S-UNI	0.05	0.1	0.2	0.3	0.4	0.5
maxSRMd2	0.4168±0.0024	0.3652±0.0008	0.2919±0.0023	0.2374±0.0023	0.1917±0.0042	0.1569±0.0035
Square root	0.4177±0.0033	0.3588±0.0025	0.2851±0.0034	0.2276±0.0021	0.1785±0.0033	0.1433±0.0026
exp-Hellinger	0.4178±0.0020	0.3608±0.0033	0.2803±0.0027	0.2181±0.0028	0.1720±0.0020	0.1348±0.0025
HILL						
maxSRMd2	0.4246±0.0040	0.3742±0.0022	0.3105±0.0033	0.2580±0.0033	0.2196±0.0039	0.1815±0.0033
Square root	0.4188±0.0030	0.3669±0.0032	0.3007±0.0025	0.2512±0.0036	0.2116±0.0026	0.1736±0.0030
exp-Hellinger	0.4191±0.0022	0.3653±0.0024	0.2974±0.0028	0.2451±0.0024	0.2004±0.0019	0.1649±0.0031
MiPOD						
maxSRMd2	0.4427±0.0026	0.3949±0.0031	0.3246±0.0034	0.2709±0.0027	0.2272±0.0037	0.1865±0.0029
Square root	0.4401±0.0028	0.3926±0.0047	0.3185±0.0022	0.2635±0.0027	0.2209±0.0036	0.1818±0.0022
exp-Hellinger	0.4426±0.0032	0.3911±0.0038	0.3148±0.0026	0.2568±0.0024	0.2104±0.0028	0.1720±0.0031
WOW						
maxSRMd2	0.3574±0.0024	0.2984±0.0020	0.2331±0.0018	0.1907±0.0028	0.1559±0.0024	0.1279±0.0030
Square root	0.3492±0.0021	0.2854±0.0033	0.2140±0.0031	0.1702±0.0026	0.1375±0.0020	0.1118±0.0033
exp-Hellinger	0.3470±0.0024	0.2820±0.0024	0.2094±0.0025	0.1645±0.0031	0.1310±0.0028	0.1068±0.0032

Table II
DETECTION ERROR $\overline{P}_{\mathrm{E}}$ FOR FOUR STEGANOGRAPHIC SCHEMES AND A RANGE OF PAYLOADS IN BITS PER PIXEL PER COLOR CHANNEL ON COLOR VERSION OF BOSSBASE WITH FLD-ENSEMBLE TRAINED WITH MAXSCRMQ1 FEATURES, THEIR SQUARE ROOTED FORM, AND TRANSFORMED USING EXPONENTIAL HELLINGER KERNEL BY SUBMODELS.

	Payload (bpp per channel)					
S-UNIWARD	0.05	0.1	0.2	0.3	0.4	0.5
SCRM	0.4549±0.0022	0.3939±0.0026	0.2977±0.0027	0.2216±0.0016	0.1710±0.0032	0.1306±0.0038
Square root	0.4499±0.0028	0.3853±0.0029	0.2885±0.0023	0.2154±0.0017	0.1630±0.0027	0.1230±0.0032
exp-Hellinger	0.4487±0.0047	0.3789±0.0031	0.2761±0.0037	0.2016±0.0017	0.1461±0.0033	0.1067±0.0028
maxSCRM	0.3866±0.0021	0.3300±0.0039	0.2480±0.0043	0.1974±0.0029	0.1561±0.0026	0.1262±0.0021
exp-H (maxSCRM)	0.3755±0.0026	0.3109±0.0024	0.2209±0.0036	0.1665±0.0029	0.1243±0.0025	0.0960±0.0021
HILL						
SCRM	0.4699±0.0024	0.4227±0.0029	0.3288±0.0022	0.2528±0.0017	0.1967±0.0024	0.1558±0.0043
Square root	0.4586±0.0035	0.4021±0.0031	0.3130±0.0022	0.2416±0.0036	0.1896±0.0025	0.1497±0.0022
exp-Hellinger	0.4520±0.0032	0.3904±0.0044	0.2927±0.0026	0.2212±0.0031	0.1724±0.0027	0.1332±0.0022
maxSCRM	0.3850±0.0036	0.3297±0.0021	0.2583±0.0028	0.2046±0.0030	0.1687±0.0049	0.1346±0.0024
exp-H (maxSCRM)	0.3732±0.0036	0.3094±0.0022	0.2343±0.0019	0.1765±0.0025	0.1398±0.0034	0.1108±0.0029
MiPOD						
SCRM	0.4557±0.0029	0.4034±0.0029	0.3081±0.0031	0.2397±0.0042	0.1872±0.0045	0.1476±0.0026
Square root	0.4477±0.0022	0.3904±0.0021	0.3006±0.0018	0.2317±0.0028	0.1812±0.0029	0.1439±0.0030
exp-Hellinger	0.4485±0.0031	0.3802±0.0032	0.2839±0.0014	0.2133±0.0034	0.1633±0.0040	0.1253±0.0034
MaxSCRM	0.4315±0.0027	0.3677±0.0023	0.2815±0.0030	0.2187±0.0029	0.1747±0.0033	0.1385±0.0040
exp-H (maxSCRM)	0.4278±0.0035	0.3562±0.0026	0.2612±0.0036	0.1955±0.0023	0.1514±0.0036	0.1179±0.0042
WOW						
SCRM	0.4507±0.0009	0.3975±0.0033	0.2997±0.0033	0.2283±0.0021	0.1793±0.0046	0.1365±0.0036
Square root	0.4367±0.0040	0.3700±0.0042	0.2750±0.0029	0.2092±0.0021	0.1641±0.0011	0.1263±0.0027
exp-Hellinger	0.4296±0.0033	0.3600±0.0029	0.2618±0.0020	0.1936±0.0022	0.1468±0.0019	0.1129±0.0018
maxSCRM	0.3183±0.0017	0.2622±0.0035	0.1993±0.0024	0.1607±0.0029	0.1347±0.0031	0.1091±0.0029
exp-H (maxSCRM)	0.2960±0.0031	0.2331±0.0045	0.1625±0.0027	0.1226±0.0034	0.0975±0.0033	0.0752±0.0022

VIII. RICH MODEL COMPACTIFICATION

Decreasing the dimensionality of the descriptors (features) may be desirable for numerous resons. For example, a more compact feature set may allow using more complex (non-linear) classifiers. Lower-dimensional features also seem essential for constructing unsupervised detectors where high-dimensional models could not be applied [45]. Finally, smaller feature dimension will decrease computational complexity of constructing the detector and evaluating it on images, which may be important for practical applications.

The topic of compressing rich models has been already investigated before within a wide variety of application scenarios. Feature filtering aimed at removing non-influential features has been shown to improve the performance of the FLD-ensemble with rich features in the presence of the cover source mismatch and a small learning database [46]. Post-selection of features combined with boosting by regression [47] has been shown to improve the detection performance of the FLD-ensemble with HOLMES features [48]. Calibrated least squares (CLS) [45] was proposed as a general method to jointly

Table III
MEAN ABSOLUTE AND MEAN SQUARE ERROR FOR LINEAR REGRESSION AND REGRESSION TREES FOR FOUR STEGANOGRAPHIC SCHEMES WITH SRMQ1, ITS SQUARE-ROOTED VERSION, AND AFTER TRANSFORMATION WITH EXPONENTIAL HELLINGER KERNEL ON BOSSBASE.

	S-UNIWARD		WOW		HILL		MiPOD	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear regression								
SRMQ1	.1683±.0010	.0454±.0008	.1778±.0011	.0507±.0009	.1809±.0011	.0522±.0013	.1803±.0012	.0524±.0013
Square root	.1540±.0006	.0391±.0007	.1640±.0008	.0447±.0011	.1698±.0007	.0472±.0015	.1696±.0012	.0475±.0013
expH-SRMQ1	.1436±.0011	.0356±.0009	.1585±.0011	.0432±.0012	.1615±.0012	.0446±.0015	.1633±.0012	.0455±.0014
Regression tree								
SRMQ1	.1347±.0010	.0324±.0004	.1478±.0011	.0377±.0004	.1546±.0013	.0406±.0005	.1539±.0011	.0403±.0004
Square root	.1251±.0009	.0291±.0005	.1394±.0010	.0350±.0004	.1456±.0008	.0375±.0005	.1469±.0014	.0381±.0005
expH-SRMQ1	.1224±.0013	.0280±.0005	.1407±.0015	.0353±.0007	.1471±.0013	.0380±.0006	.1492±.0011	.0388±.0005

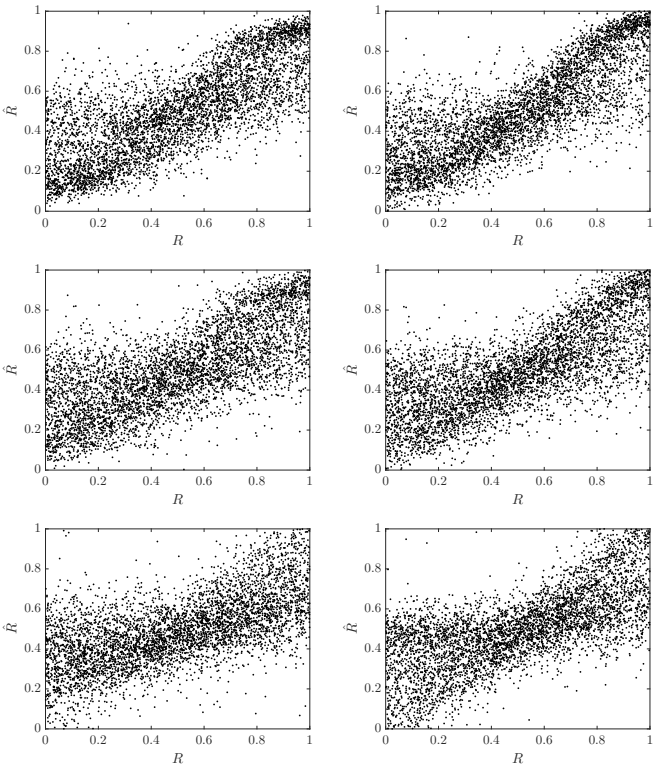

Figure 3. Selected scatter plots showing the estimated payload $\hat{R}$ in bpp vs. the true payload R for S-UNIWARD, HILL, and MiPOD (top to bottom) with the SRMQ1 feature set (left column) and the same set transformed with exponential Hellinger kernel (right column). The plots for S-UNIWARD and HILL are for regression trees while a linear regressor is shown for MiPOD.

compactify rich models that minimizes the scatter of cover features while maximizing the separation between cover and stego features. Random projections for compactification was investigated in [49].

In this section, we describe a different feature compactification method that is unaware of the distribution of the stego class. Since it is driven solely by the distribution of cover features, it may be used for building universal (unsupervised) steganography detectors. At this point, we would like to state that the above cited feature selection / filtering methods can likely be combined with the proposed non-linear feature transformation to provide a possible additional boost and/or to further compress the feature dimensionality. Since the best compactification of rich models is likely application-dependent due to various inevitable trade offs between computational complexity, detection accuracy, robustness to source, and generalization to unseen embedding, in the interest of keeping the findings of this paper on a more general level and balanced, we refrain from investigating these enticing research topics in this paper, leaving them for potential future work.

The non-linear transform as explained in Section II can be easily adapted for reducing the dimensionality of the feature vector by simply retaining the first E coordinates in $\hat{\varphi}$, $a = 1, \dots, E$, corresponding to the largest eigenvalues λ_a. This is similar in spirit to applying a regular PCA to cover features. Since the compactification only depends on the cover source, it is potentially useful for unsupervised universal steganalysis.

Even though the detection errors of individual SRM submodels increase with the decreased number of retained coordinates, the entire rich model may still perform rather well when compacted because the submodels "compensate for each other weaknesses." This is confirmed in Figure 4, which shows the detection error $\overline{P}_{\mathrm{E}}$ as a function of the number of retained coordinates for binary detectors trained on BOSSbase images (left) and BOSSColor (right). Even when retaining only 10% of the coordinates, $E = 0.1 \times D$ for each submodel of dimension D, there still appears to be a small gain in detection accuracy w.r.t. the original maxSRMd2 feature.

Surprisingly, for quantitative detectors the dimensionality reduction further increases their accuracy. As Figure 5 shows, by retaining only 40% of the rich feature, the MAE further decreases to a combined improvement of 18–28% (from 13–22% for full dimensionality) for linear regressors and 8.4–17% (from 4-14% for full dimensionality), depending on the embedding algorithm. The MSE followed similar trends.

IX. Discussion

The degree of improvement due to the non-linear mapping will necessarily depend on two factors – the non-linearity of the decision boundary and the accuracy of the Nyström approximation. By definition, no gain will be observed when the decision boundary is (near) linear. Small or no gain may also be due to the poor accuracy of

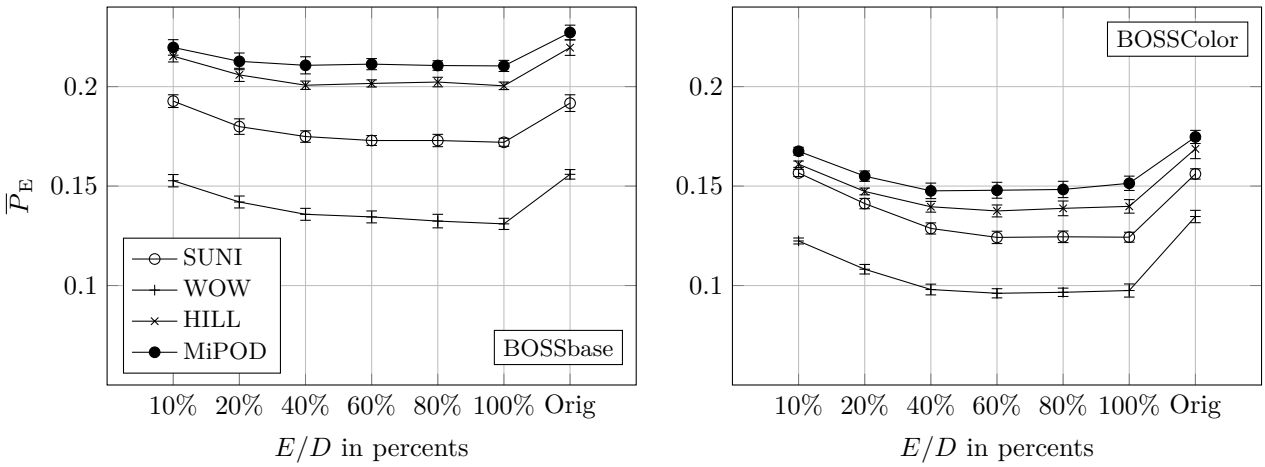

Figure 4. Detection error $\overline{P}_{\mathrm{E}}$ as a function of the relative number of retained coordinates, E/D. Tested payload 0.4 bpp, exp-Hellinger kernel. Left: BOSSbase (maxSRMd2), Right: BOSSColor (maxSCRM).

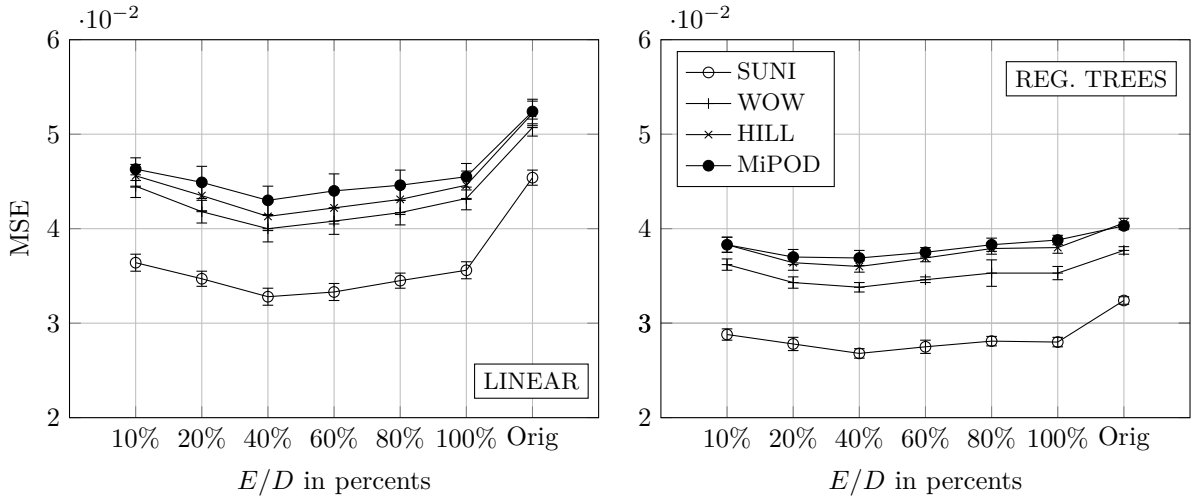

Figure 5. MSE as a function of the relative number of retained coordinates, E/D. BOSSbase, SRMQ1 feature set. Left: OLS, Right: regression tree. Note the gain when retaining 40% of feature dimensionality.

the Nyström approximation to approximate the infinite-dimensional mapping (9)–(10).

In particular, as reported in [31], [32], for the steganographic algorithm nsF5 [50], [51], the decision boundary in the JPEG rich model (JRM) [31] and CC-C300 [32] is (near) linear because the features are formed by two-dimensional co-occurrences of DCT coefficients directly affected by the embedding. Consequently, the proposed non-linear maps did not provide any boost when applied to these feature sets. No boost of non-linear transforms was observed for the projection SRM (PSRM) model [9] for spatial domain steganography and for JPEG-phase-aware rich models [16]–[19] when detecting JPEG steganographic algorithms J-UNIWARD [4] and UED [3], [6]. We note that the non-linearly transformed SRM offers better detection than the PSRM with a much lower computational complexity.

Finally, we wish to comment on the possibility to combine the proposed transformation with normalization techniques. It is customary in machine learning to normalize the features by, e.g., scaling each bin to be within a unit interval or to scale it to zero mean and unit variance before applying a given machine learning tool. Such scaling cannot affect detection accuracy with the FLD-ensemble because the individual base learners are built using the FLD, which is oblivious to linear transformations. By the same token, linear transformations, such as PCA, will have no effect on detection with the FLD-ensemble. The same is true for regression trees with "base estimators" constructed as ordinary least square regressors.

On the other hand, non-linear normalization can be combined with the proposed approach and may potentially provide further boost. In particular, the authors tested whether the non-linear normalization called random conditioning proposed in [52] could provide detection boost in combination with the non-linear mapping. Since these two feature preprocessing operations do not commute, both orders were tested on various scenarios of binary classification of grayscale images with the SRMQ1 model. Neither, however, lead to any statistically significant improvement.

Random conditioning before transformation is expected to have no or little effect because the proposed non-linear transformation with the Hellinger kernel executes L_1 normalization of features, which is equivalent to conditioning on each submodel.

X. Conclusions

Currently, training a kernelized support vector machine with high dimensional rich representations on tens of thousands of images is computationally infeasible for research in steganalysis considering that the established practice calls for reporting detection performance on multiple splits of the image source into training and testing sets and considering the need to show performance across payloads and multiple embedding schemes. The computational complexity is even higher for quantitative detectors that estimate the payload size (support vector regressors) because the search for the hyperparameters is three-dimensional. To deal with this complexity, the community resorted to simpler machine learning paradigms, such as linear classifiers and regression trees.

In this paper, we describe a simple method for increasing the performance of the detectors by transforming the features prior to training a low-complexity classifier / regressor. The methodology draws from recent advances in approximations of positive semidefinite kernels proposed in the literature. The mapping is derived from symmetrized Ali–Silvey distances (kernels) and estimated using the Nyström approximation. The approach is applied to rich models by learning the transformation separately for each submodel in order to keep the computational complexity low. A small fixed set of cover features is needed to train the transform, which only depends on a handful of cover features and not on the steganographic scheme or the embedded payload.

Coupled with the ensemble classifier, a consistent gain in detection accuracy between 2–4% was observed for binary classifiers with the selection-channel-aware maxSRMd2 features for grayscale images as well as the Spatio-Color Rich Model for steganalysis of color images. For quantitative detectors (payload regressors), the gain in terms of decreased statistical spread of the payload size estimate measured as MSE estimate was in the range of 18-28% for linear regressors and 8–17% for regression trees.

The proposed methodology naturally lends itself for unsupervised dimensionality reduction by simply retaining fewer transformed dimensions. In particular, it is possible to compactify the rich descriptor by a factor of 10 without losing the detection performance of the original (non-transformed) feature vector. For quantitative detectors, the features can be compacted by 60% while further decreasing the statistical spread of the payload estimates. This dimensionality reduction could be useful for unsupervised universal steganalysis detectors.

The code for all algorithms (steganographic methods, feature extractors, and classifiers) is available for download from http://dde.binghamton.edu/download/.

References

[1] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Information Hiding, 12th International Conference* (R. Böhme and R. Safavi-Naini, eds.), vol. 6387 of Lecture Notes in Computer Science, (Calgary, Canada), pp. 161–177, Springer-Verlag, New York, June 28–30, 2010.

[2] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Fourth IEEE International Workshop on Information Forensics and Security*, (Tenerife, Spain), December 2–5, 2012.

[3] L. Guo, J. Ni, and Y.-Q. Shi, "An efficient JPEG steganographic scheme using uniform embedding," in *Fourth IEEE International Workshop on Information Forensics and Security*, (Tenerife, Spain), December 2–5, 2012.

[4] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion design for steganography in an arbitrary domain," *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, vol. 2014:1, 2014.

[5] B. Li, M. Wang, and J. Huang, "A new cost function for spatial image steganography," in *Proceedings IEEE, International Conference on Image Processing, ICIP*, (Paris, France), October 27–30, 2014.

[6] L. Guo, J. Ni, and Y. Q. Shi, "Uniform embedding for efficient JPEG steganography," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 5, pp. 814–825, 2014.

[7] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, 2016.

[8] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 868–882, June 2011.

[9] V. Holub and J. Fridrich, "Random projections of residuals for digital image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 1996–2006, December 2013.

[10] L. Chen, Y. Shi, P. Sutthiwan, and X. Niu, "A novel mapping scheme for steganalysis," in *International Workshop on Digital Forensics and Watermaking* (Y. Shi, H.-J. Kim, and F. Perez-Gonzalez, eds.), vol. 7809 of *LNCS*, pp. 19–33, Springer Berlin Heidelberg, 2013.

[11] L. Chen, Y.-Q. Shi, and P. Sutthiwan, "Variable multi-dimensional co-occurrence for steganalysis," in *Digital Forensics and Watermarking, 13th International Workshop, IWDW*, vol. 9023, (Taipei, Taiwan), pp. 559–573, Springer, October 1–4 2014.

[12] W. Tang, H. Li, W. Luo, and J. Huang, "Adaptive steganalysis against WOW embedding algorithm," in *2nd ACM IH&MMSec. Workshop* (A. Uhl, S. Katzenbeisser, R. Kwitt, and A. Piva, eds.), (Salzburg, Austria), pp. 91–96, June 11–13, 2014.

[13] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *IEEE International Workshop on Information Forensics and Security*, (Atlanta, GA), December 3–5, 2014.

[14] T. Denemark and J. Fridrich, "Improving selection-channel-aware steganalysis features," in *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2016* (A. Alattar and N. D. Memon, eds.), (San Francisco, CA), February 14–18, 2016.

[15] W. Tang, H. Li, W. Luo, and J. Huang, "Adaptive steganalysis based on embedding probabilities of pixels," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 734–745, April 2016.

[16] T. Denemark, M. Boroumand, and J. Fridrich, "Steganalysis features for content-adaptive JPEG steganography," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 1736–1746, June 2016.

[17] V. Holub and J. Fridrich, "Low-complexity features for JPEG steganalysis using undecimated DCT," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 219–228, Feb 2015.

[18] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang, "Steganalysis of adaptive JPEG steganography using 2D Gabor filters," in

3rd ACM IH&MMSec. Workshop (P. C. na, J. Fridrich, and A. Alattar, eds.), (Portland, Oregon), June 17–19, 2015.

[19] V. Holub and J. Fridrich, "Phase-aware projection model for steganalysis of JPEG images," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2015* (A. Alattar and N. D. Memon, eds.), vol. 9409, (San Francisco, CA), February 8–12, 2015.

[20] M. Goljan, R. Cogranne, and J. Fridrich, "Rich model for steganalysis of color images," in *Sixth IEEE International Workshop on Information Forensics and Security*, (Atlanta, GA), December 3–5, 2014.

[21] H. Abdulrahman, M. Chaumont, P. Montesinos, and B. Magnier, "Color image steganalysis using RGB channel geometric transformation measures," *Wiley Journal on Security and Communication Networks*, February 2016.

[22] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2015* (A. Alattar and N. D. Memon, eds.), vol. 9409, (San Francisco, CA), February 8–12, 2015.

[23] G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, pp. 708–712, May 2016.

[24] T. Pevný, J. Fridrich, and A. D. Ker, "From blind to quantitative steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 445–454, 2011.

[25] J. Kodovský and J. Fridrich, "Quantitative steganalysis using rich models," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2013* (A. Alattar, N. D. Memon, and C. Heitzenrater, eds.), vol. 8665, (San Francisco, CA), pp. O 1–11, February 5–7, 2013.

[26] J. Kodovský, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.

[27] R. Cogranne and J. Fridrich, "Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 2627–2642, December 2015.

[28] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich, "Is ensemble classifier needed for steganalysis in high-dimensional feature spaces?," in *IEEE International Workshop on Information Forensics and Security*, (Rome, Italy), November 16–19, 2015.

[29] D. C.-L. Fong and M. Saunders, "LSMR: An iterative algorithm for sparse least-squares problems," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2950–2971, 2011.

[30] I. Lubenko and A. D. Ker, "Steganalysis with mismatched covers: Do simple classifiers help," in *Proc. 13th ACM Workshop on Multimedia and Security* (J. Dittmann, S. Katzenbeisser, and S. Craver, eds.), (Coventry, UK), pp. 11–18, September 6–7 2012.

[31] J. Kodovský and J. Fridrich, "Steganalysis of JPEG images using rich models," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012* (A. Alattar, N. D. Memon, and E. J. Delp, eds.), vol. 8303, (San Francisco, CA), pp. 0A 1–13, January 23–26, 2012.

[32] J. Kodovský and J. Fridrich, "Steganalysis in high dimensions: Fusing classifiers built on random subspaces," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics III* (A. Alattar, N. D. Memon, E. J. Delp, and J. Dittmann, eds.), vol. 7880, (San Francisco, CA), pp. OL 1–13, January 23–26, 2011.

[33] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 480–492, March 2012.

[34] F. Perronnin, J. Sanchez, and Y. Liu, "Large-scale image categorization with explicit data embedding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2297–2304, June 2010.

[35] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2911–2918, June 2012.

[36] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: a new blind image splicing detector," in *IEEE International Workshop on Information Forensics and Security*, (Rome, Italy), November 16–19, 2015.

[37] M. Boroumand and J. Fridrich, "Boosting steganalysis with explicit feature maps," in *4th ACM IH&MMSec. Workshop* (F. Perez-Gonzales, F. Cayre, and P. Bas, eds.), (Vigo, Spain), June 20–22, 2016.

[38] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society Series B*, vol. 28, pp. 131–142, 1966.

[39] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

[40] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 15, no. 5, pp. 1155–1178, 2007.

[41] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems (NIPS)* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 1177–1184, Curran Associates, Inc., 2008.

[42] T. Filler, T. Pevný, and P. Bas, "BOSS (Break Our Steganography System)." http://www.agents.cz/boss, July 2010.

[43] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *Advances in Neural Information Processing Systems (NIPS) 22* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 135–143, Curran Associates, Inc., 2009.

[44] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Advances in Neural Information Processing Systems (NIPS) 22* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1509–1517, Curran Associates, Inc., 2009.

[45] T. Pevný and A. D. Ker, "The challenges of rich features in universal steganalysis," in *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2013* (A. Alattar, N. D. Memon, and C. Heitzenrater, eds.), vol. 8665, (San Francisco, CA), pp. 0M 1–15, February 5–7, 2013.

[46] J. Pasquet, S. Bringay, and M. Chaumont, "Steganalysis with cover-source mismatch and a small learning database," in *22nd European Signal Processing Conference (EUSIPCO)*, pp. 2425–2429, 2014.

[47] M. Chaumont and S. Kouider, "Steganalysis by ensemble classifiers with boosting by regression, and postselection of features," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1133–1136, September 2012.

[48] J. Fridrich, J. Kodovský, M. Goljan, and V. Holub, "Steganalysis of content-adaptive steganography in spatial domain," in *Information Hiding, 13th International Conference* (T. Filler, T. Pevný, A. Ker, and S. Craver, eds.), Lecture Notes in Computer Science, (Prague, Czech Republic), pp. 102–117, May 18–20, 2011.

[49] P. Wang, Z. Wei, and L. Xiao, "Fast projections of spatial rich model feature for digital image steganalysis," *Soft Computing*, pp. 1–9, 2016.

[50] J. Fridrich, T. Pevný, and J. Kodovský, "Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities," in *Proceedings of the 9th ACM Multimedia & Security Workshop* (J. Dittmann and J. Fridrich, eds.), (Dallas, TX), pp. 3–14, September 20–21, 2007.

[51] A. Westfeld, "High capacity despite better steganalysis (F5 – a steganographic algorithm)," in *Information Hiding, 4th International Workshop* (I. S. Moskowitz, ed.), vol. 2137 of Lecture Notes in Computer Science, (Pittsburgh, PA), pp. 289–302, Springer-Verlag, New York, April 25–27, 2001.

[52] M. Boroumand and J. Fridrich, "Non-linear feature normalization in steganalysis," in *5th ACM IH&MMSec. Workshop* (M. Stamm, M. Kirchner, and S. Voloshynovskiy, eds.), (Philadelphia, PA), June 20–22, 2017.