LA-UR-15-25934 (Accepted Manuscript)

Problematic Projection to the In-Sample Subspace for a Kernelized Anomaly Detector

Theiler, James Patrick
Grosklos, Guenchik Jun

To be published in: IEEE Geoscience and Remote Sensing Letters

DOI to publisher's version: 10.1109/LGRS.2016.2516985

Permalink to record: http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-15-25934

Problematic projection to the in-sample subspace for a kernelized anomaly detector

James Theiler and Guen Grosklos
Intelligence and Space Research Division,
Los Alamos National Laboratory, Los Alamos, NM 87545, USA

***Abstract*—We examine the properties and performance of kernelized anomaly detectors, with an emphasis on the Mahalanobis-distance based kernel-RX (KRX) algorithm. Although the detector generally performs well for high-bandwidth Gaussian kernels, it exhibits problematic (in some cases catastrophic) performance for distances that are large compared to the bandwidth. By comparing KRX to two other anomaly detectors, we can trace the problem to a projection in feature space, which arises when a pseudoinverse is used on the covariance matrix in that feature space. We show that a regularized variant of KRX overcomes this difficulty and achieves superior performance over a wide range of bandwidths.**

***Index Terms*—Adaptive signal detection, algorithms, covariance matrices, data models, detectors, multidimensional signal processing, pattern recognition, remote sensing. singular value decomposition, spectral analysis.**

I. Introduction

Anomaly detection is the unsupervised identification of data samples (*e.g.*, pixels in a hyperspectral image) that are unusual with respect to the rest of the data [1], [2]. For instance, if the data are modeled by a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance C, then the Mahalanobis distance [3]

$$\mathcal{A}(\mathbf{r}) = (\mathbf{r} - \boldsymbol{\mu})^T C^{-1} (\mathbf{r} - \boldsymbol{\mu}) \tag{1}$$

provides a simple measure of how anomalous the point $\mathbf{r}$ is; this measure monotonically increases for decreasing likelihood that $\mathbf{r}$ is drawn from the distribution. The use of Mahalanobis distance for multispectral and hyperspectral anomaly detection was popularized by Reed and Yu [4] and is commonly referred to as the RX algorithm. A kernelization of the Mahalanobis distance was proposed by Cremers *et al.* [5] and adopted for hyperspectral anomaly detection by Kwon and Nasrabadi [6]. In this approach, a data sample is mapped to a feature space, and Mahalanobis distance is computed in that feature space. Where RX effectively assumes that the data samples are drawn from an elliptically-contoured distribution, kernel-RX (KRX) can accommodate more convoluted contours.

In this paper, we identify a property of KRX – an implicit projection in feature space – that leads to diminished performance, particularly at small bandwidths, and we show that a simple regularization scheme alleviates the problem. Section II derives a family of four kernelized anomaly detectors, two of which employ a projection to the in-sample subspace and two of which do not. These derivations clarify the role of this projection in kernelized anomaly detectors. Section III deploys these anomaly detectors first on simple one- and two-dimensional problems (where the distinctions between the detectors are pronounced and conspicuous), and then on real hyperspectral image data. In Section IV, we briefly conclude.

II. Kernelized anomaly detection

In this section we derive KRX, and a regularized variant KRX-reg. But before we do that, we derive two other kernelized anomaly detectors, one that is standard (KDE, or kernel density estimation [7], [8]) and one that is new (KDE-flat). Although the KDE-flat detector is our own invention, we do not advocate its use in practice because its performance is poor. We introduce this detector to illuminate the problem caused by projection to the data-defined subspace of feature space. This is a problem that it shares with KRX.

A. Notation

The background dataset is comprised of N samples (pixels, usually, in hyperspectral applications), with each sample a d-dimensional point: $\mathbf{x}_n \in \mathcal{X} \subseteq \mathbb{R}^d$ for $1 \le n \le N$. Our aim is to estimate a function $\mathcal{A} : \mathcal{X} \to \mathbb{R}$ that characterizes the relative anomalousness of points in the data space $\mathcal{X}$.

If the data are drawn from a probability density function $p(\mathbf{x})$, then anomalies occur where $p(\mathbf{x})$ is small, so what we seek from an anomaly detector $\mathcal{A}(\mathbf{x})$ is some negative monotonic function of $p(\mathbf{x})$.

For kernel-based methods, the data samples are mapped to a feature space $\mathcal{F}$ by a function Φ. That is $\Phi : \mathcal{X} \to \mathcal{F}$, and in particular, $\Phi(\mathbf{r}) \in \mathcal{F}$ is the mapping of the point $\mathbf{r} \in \mathcal{X}$ into feature space. This feature space has the property that dot products can be expressed as a scalar function of points in the original data space: $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$; more specifically,

$$k(\mathbf{r}, \mathbf{s}) = \Phi(\mathbf{r})^T \Phi(\mathbf{s}) \in \mathbb{R}. \tag{2}$$

The "kernel trick" is the recognition that, by specifying the kernel function $k(\mathbf{r}, \mathbf{s})$, one may not need to explicitly evaluate $\Phi(\mathbf{r})$ or $\Phi(\mathbf{s})$. A popular choice, and the one we consider here, is the Gaussian radial basis function kernel:

$$k(\mathbf{r}, \mathbf{s}) = \exp(-\|\mathbf{r} - \mathbf{s}\|^2 / 2\sigma^2), \tag{3}$$

where the parameter σ is called the bandwidth.

It is useful to consider the centroid of the data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in the feature space, $\boldsymbol{\mu}_\Phi = \frac{1}{N} \sum_{n=1}^{N} \Phi(\mathbf{x}_n)$, and in terms of this centroid, to define a centered feature map

$\Phi_c(\mathbf{r}) = \Phi(\mathbf{r}) - \boldsymbol{\mu}_\Phi$. This new feature map can then be used to define a centered kernel function:

$$\begin{aligned} k_c(\mathbf{r},\mathbf{s}) &= \Phi_c(\mathbf{r})^T \Phi_c(\mathbf{s}) \\ &= k(\mathbf{r},\mathbf{s}) - \frac{1}{N}\sum_n k(\mathbf{r},\mathbf{x}_n) - \frac{1}{N}\sum_m k(\mathbf{x}_m,\mathbf{s}) \\ &\quad + \frac{1}{N^2}\sum_{n,m} k(\mathbf{x}_n,\mathbf{x}_m). \end{aligned} \tag{4}$$

B. Kernel density estimation (KDE)

In traditional kernel density estimation [8] (also called Parzen windows [7]), a probability density $p(\mathbf{r})$ is estimated in terms of the data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

$$\hat{p}(\mathbf{r}) = \frac{1}{N}\sum_n c_0^{-1} \exp(-\|\mathbf{r}-\mathbf{x}_n\|/2\sigma^2), \tag{5}$$

where $c_0 = (2\pi\sigma^2)^{d/2}$ is a constant that normalizes the density function. Note that $\hat{p}(\mathbf{r})$ is a *fixed* kernel density estimator; there is a large family of *variable* kernel density estimators, for which the kernel itself is different for different points $\mathbf{r}$; *e.g.*, [8], [9], [10], [11].

Following an appendix in Cremers *et al.* [5], we derive the KDE detector as Euclidean distance in feature space.

$$\mathcal{A}_{\text{KDE}}(\mathbf{r}) = \Phi_c(\mathbf{r})^T \Phi_c(\mathbf{r}) = k_c(\mathbf{r},\mathbf{r}), \tag{6}$$

where $\Phi_c(\mathbf{r})$ is the centered feature map and k_c is the centered kernel function. From Eq. (4), we can write the KDE detector

$$\mathcal{A}_{\text{KDE}}(\mathbf{r}) = k(\mathbf{r},\mathbf{r}) - \frac{2}{N}\sum_n k(\mathbf{r},\mathbf{x}_n) + \frac{1}{N^2}\sum_{n,m} k(\mathbf{x}_n,\mathbf{x}_m). \tag{7}$$

The third term is a constant, and for a radial-basis kernel $k(\mathbf{r},\mathbf{r})$ is also a constant, so

$$\mathcal{A}_{\text{KDE}}(\mathbf{r}) = c_1 - \frac{2}{N}\sum_n k(\mathbf{r},\mathbf{x}_n) = c_1 - 2c_0\,\hat{p}(\mathbf{r}), \tag{8}$$

with c_0 and c_1 constants. Thus, the KDE anomaly detector is a negative monotonic function of the probability density $\hat{p}(\mathbf{r})$ that was estimated using kernel density estimation.

C. Flattened KDE (KDE-flat)

In this subsection, we derive a variant of KDE that includes a "flattening" of the input; a projection to the subspace of $\mathcal{F}$ spanned by the training data. As we will see in Section III, the effect of this projection is significant.

To begin, define the data matrix in centered feature space:

$$\mathbf{X}_\Phi = \left[\Phi_c(\mathbf{x}_1)\cdots\Phi_c(\mathbf{x}_N)\right], \tag{9}$$

Let r be the rank of this matrix (observe that $r \le N-1$ since the centroid has been subtracted). Express $\mathbf{X}_\Phi$ with a singular value decomposition

$$\mathbf{X}_\Phi = V_\Phi \Lambda^{1/2} W^T. \tag{10}$$

Here V_Φ is an orthogonal matrix with r columns (so $V_\Phi^T V_\Phi = I$), Λ is a diagonal $r \times r$ matrix with positive entries, and W is an orthogonal $N \times r$ matrix (for which $W^T W = I$).

Note that columns of V_Φ are eigenvectors of the covariance matrix $C_\Phi = \mathbf{X}_\Phi \mathbf{X}_\Phi^T$, and columns of W are eigenvectors of the centered Gram matrix

$$K_c = \mathbf{X}_\Phi^T \mathbf{X}_\Phi = \begin{bmatrix} k_c(\mathbf{x}_1,\mathbf{x}_1) & \cdots & k_c(\mathbf{x}_1,\mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k_c(\mathbf{x}_N,\mathbf{x}_1) & \cdots & k_c(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix}. \tag{11}$$

Also, the diagonal elements of Λ are the positive eigenvalues of both C_Φ and K_c.

Note further that V_Φ projects (flattens) vectors in the feature space into an r-dimensional subspace. We can thus modify our KDE anomaly detector by computing Euclidean distance in this subspace (see Eq. 1 in Nasrabadi [12]):

$$\mathcal{A}_{\text{flat}}(\mathbf{r}) = \Phi_c(\mathbf{r})^T V_\Phi V_\Phi^T \Phi_c(\mathbf{r}). \tag{12}$$

From Eq. (10), we can write $V_\Phi = X_\Phi W \Lambda^{-1/2}$, and so

$$\begin{aligned} \mathcal{A}_{\text{flat}}(\mathbf{r}) &= \Phi_c(\mathbf{r})^T \mathbf{X}_\Phi W \Lambda^{-1} W^T \mathbf{X}_\Phi^T \Phi_c(\mathbf{r}) \\ &= Z_c(\mathbf{r})^T K_c^{-1} Z_c(\mathbf{r}), \end{aligned} \tag{13}$$

where $Z_c(\mathbf{r})$ can be expressed in terms of the centered kernel:

$$Z_c(\mathbf{r}) = \mathbf{X}_\Phi^T \Phi_c(\mathbf{r}) = \begin{bmatrix} k_c(\mathbf{x}_1,\mathbf{r}) \\ \vdots \\ k_c(\mathbf{x}_N,\mathbf{r}) \end{bmatrix}. \tag{14}$$

D. Kernel-RX (KRX)

The KRX idea is to use a Mahalanobis distance instead of a Euclidean distance in the feature space. That is:

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \Phi_c(\mathbf{r})^T C_\Phi^{-1} \Phi_c(\mathbf{r}), \tag{15}$$

where the covariance matrix C_Φ is determined from the data in feature space:

$$C_\Phi = \sum_n \Phi_c(\mathbf{r})\Phi_c(\mathbf{r})^T = X_\Phi X_\Phi^T = V_\Phi \Lambda V_\Phi^T \tag{16}$$

where X_Φ was defined in Eq. (9), and decomposed in Eq. (10). The problem with KRX, as it is expressed in Eq. (15), is that C_Φ is not invertible. Although not explicitly discussed in [6], the approach taken in [6] uses the pseudoinverse. That is,

$$C_\Phi^{-1} = (V_\Phi \Lambda V_\Phi^T)^{-1} = V_\Phi \Lambda^{-1} V_\Phi^T. \tag{17}$$

The ambiguous left-hand side is simply replaced with the well-defined right-hand side. We can use $V_\Phi = X_\Phi W \Lambda^{-1/2}$, obtained from Eq. (10), to further simplify

$$\begin{aligned} C_\Phi^{-1} &= (X_\Phi W \Lambda^{-1/2}) \Lambda^{-1} (\Lambda^{-1/2} W^T \mathbf{X}_\Phi^T) \\ &= \mathbf{X}_\Phi W \Lambda^{-2} W^T \mathbf{X}_\Phi^T = \mathbf{X}_\Phi K_c^{-2} \mathbf{X}_\Phi^T, \end{aligned} \tag{18}$$

where K_c is the centered Gram matrix defined in Eq. (11), and K_c^{-2} refers to the pseudoinverse of K_c^2. Thus,

$$\begin{aligned} \mathcal{A}_{\text{KRX}}(\mathbf{r}) &= \Phi_c(\mathbf{r})^T X_\Phi K_c^{-2} \mathbf{X}_\Phi^T \Phi_c(\mathbf{r}) \\ &= Z_c(r)^T K_c^{-2} Z_c(r), \end{aligned} \tag{19}$$

where $Z_c(r) = \mathbf{X}_\Phi^T \Phi_c(\mathbf{r})$ was defined in Eq. (14).

It is important to recognize that the pseudoinverse involves the projection of $\Phi_c(\mathbf{r})$ to $V_\Phi^T \Phi_c(\mathbf{r})$. Therefore, it is more appropriate to think of KRX as the Mahalanobis variant not of KDE, but of KDE-flat.

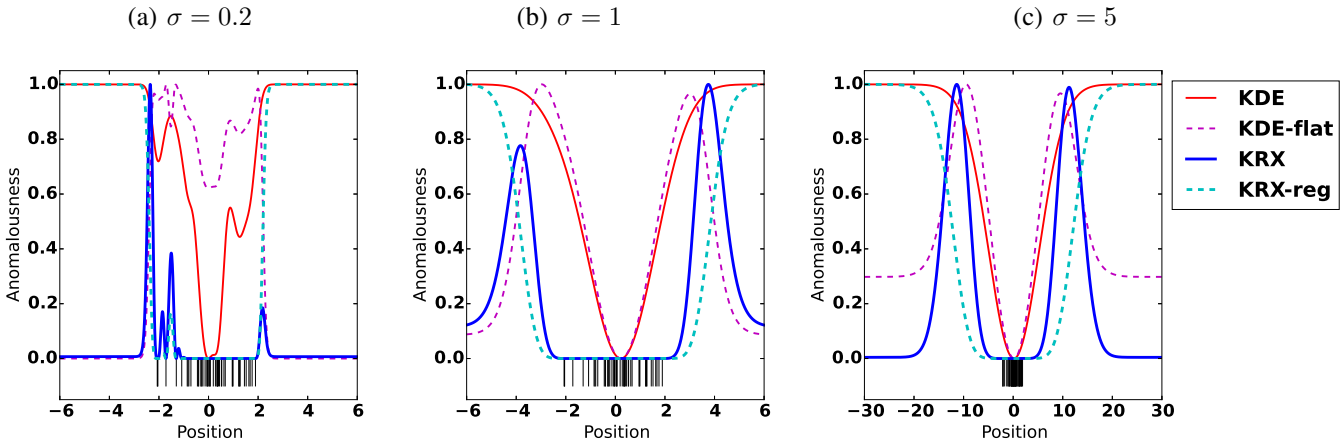

Fig. 1. In this simple one-dimensional example, $N = 50$ training points are drawn from a Gaussian distribution with zero mean and unit variance. Anomaly detectors are derived in terms of this training data, and anomalousness (scaled to range from zero to one) is plotted as a function of position. Bandwidths are chosen to be (a) $\sigma = 0.2$, (b) $\sigma = 1$, and (c) $\sigma = 5$.

E. Regularized kernel-RX (KRX-reg)

To deal with the singular matrix C_Φ in Eq. (16), KRX employs the pseudoinverse. A common alternative approach for inverting singular and near-singular matrices is to regularize them first. Thus, in place of the sample covariance defined in Eq. (16), we can employ a regularization operation:

$$C'_\Phi = C_\Phi + \lambda I \tag{20}$$

for some small λ. Since C_Φ is not of full rank it is useful to employ singular value decomposition, and write (see Eq. (16)) $C_\Phi = V_\Phi \Lambda V_\Phi^T$. Thus,

$$C'_\Phi = V_\Phi(\Lambda + \lambda)V_\Phi^T + \lambda(I - V_\Phi V_\Phi^T). \tag{21}$$

Since the two terms in the sum are orthogonal to each other, we can invert the sum by inverting the individual components:

$$C'^{-1}_\Phi = V_\Phi(\Lambda + \lambda)^{-1}V_\Phi^T + \lambda^{-1}(I - V_\Phi V_\Phi^T) \tag{22}$$

and the anomalousness is given by the Mahalanobis distance with respect to this regularized covariance matrix:

$$\begin{aligned}\mathcal{A}_{\text{reg}}(\mathbf{r}) &= \Phi_c(\mathbf{r})^T C'^{-1}_\Phi \Phi_c(\mathbf{r}) \\ &= \Phi_c(\mathbf{r})^T V_\Phi(\Lambda + \lambda)^{-1}V_\Phi^T \Phi_c(\mathbf{r}) \\ &\quad + \lambda^{-1}\Phi_c(\mathbf{r})^T(I - V_\Phi V_\Phi^T)\Phi_c(\mathbf{r}) \\ &= \mathcal{A}^*_{\text{KRX}}(\mathbf{r}) + \lambda^{-1}\left[\mathcal{A}_{\text{KDE}}(\mathbf{r}) - \mathcal{A}_{\text{flat}}(\mathbf{r})\right]\end{aligned} \tag{23}$$

where $\mathcal{A}^*_{\text{KRX}}$ is computed using

$$\mathcal{A}^*_{\text{KRX}}(\mathbf{r}) = Z_c(\mathbf{r})^T K_c^{-1/2}(K_c + \lambda I)^{-1}K_c^{-1/2}Z_c(\mathbf{r}), \tag{24}$$

and $K_c^{-1/2}$ is the matrix square root of the pseudoinverse of the Gram matrix. As a practical matter, we remark that $\mathcal{A}^*_{\text{KRX}}$ and $\mathcal{A}_{\text{KRX}}$ behave nearly identically. But $\mathcal{A}_{\text{reg}}$ is considerably different from $\mathcal{A}^*_{\text{KRX}}$ (as is indicated by the second term in Eq. (23), which, with the λ^{-1} pre-factor, is large). We note that KRX-reg does not simply regularize the covariance matrix K_c in the KRX formula, given in Eq. (19); the regularization is of C_Φ and takes place in the feature space.

For the experiments reported here, we used the small but numerically relevant value $\lambda = 10^{-8}\max_i \Lambda_{ii}$. We note that in the $\lambda \to 0$ limit, Eq. (23) is dominated by the regularization term; this term, $\Phi_c(\mathbf{r})^T(I - V_\Phi V_\Phi^T)\Phi_c(\mathbf{r})$, actually provides an anomaly detector in its own right [13], [12].

III. Comparision of Algorithms

To compare the kernel-based anomaly detectors derived in the previous section, we apply them both to artificial one- and two-dimensional examples and to real hyperspectral data.

A. One-dimensional example

We begin with a very simple one-dimensional example; N points are drawn from a standard (zero mean and unit variance) normal distribution. For this distribution, we expect anomalousness to be minimal at position $\mathbf{r} = 0$ (the peak of the probability distribution), and to increase monotonically with distance from zero. As seen in Fig. 1, this behavior is indeed observed with the KDE detector, but we see problems with KDE-flat and KRX, the two anomaly detectors that employ a projection to the in-sample subspace. Anomalousness does initally increase with increasing distance from zero, for these detectors, but then it reverses itself and gets smaller. One can try to address this problem by using a larger bandwidth, σ, but however large the bandwidth is chosen to be (*e.g.,* $\sigma = 5$ in Fig. 1(c)), there is a distance beyond which anomalousness decreases with increasing distance. This problem with KRX is fixed by KRX-reg.

B. Two-dimensional examples

With the two-dimensional examples illustrated in Fig. 2, we can see how the more adaptive KRX is able to more compactly enclose the data, but at the same time how the anomalousness decreases for distant outliers. This non-monotonicity with distance is even more pronounced for KDE-flat, but the phenomenon is present in both KDE-flat and KRX.

While Fig. 2(a,b,c,d) uses bandwidth $\sigma = 5$, the corresponding Fig. 3(a) illustrates how performance (as measured by volume enclosed by a contour) varies with σ. It is clear that KDE works best at small values of σ, but KDE-flat and

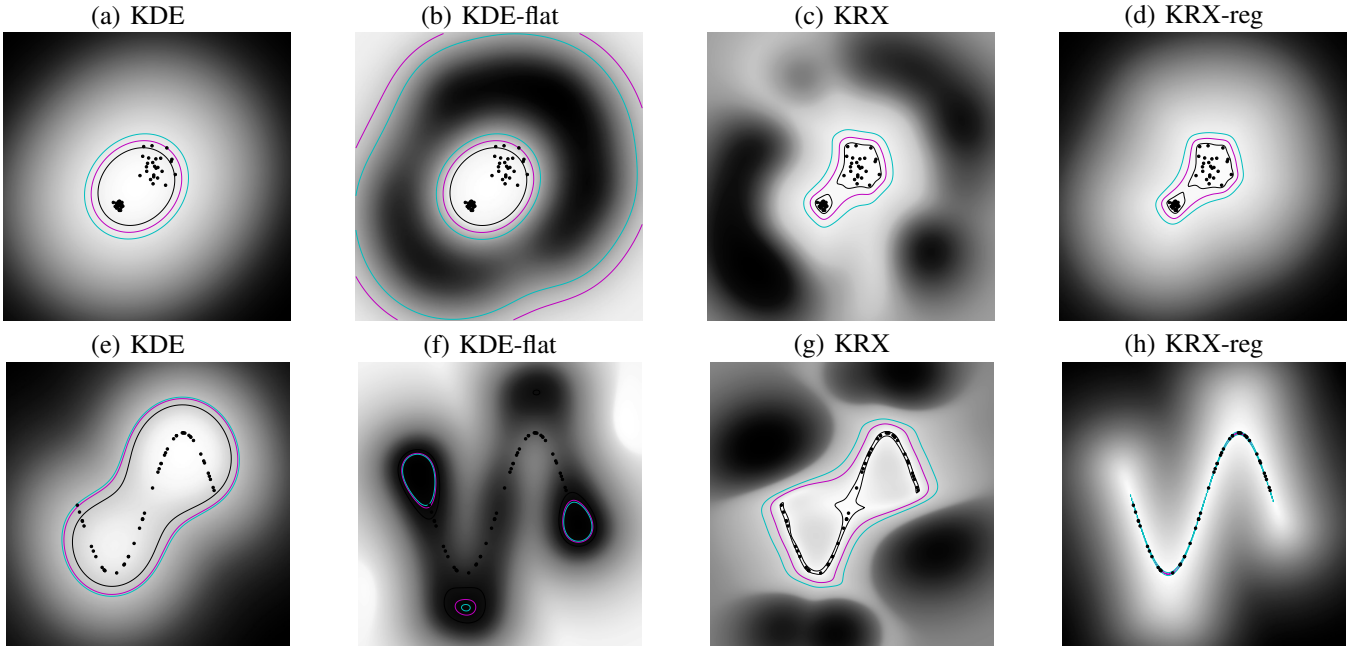

Fig. 2. In these two-dimensional examples, $N = 50$ training points are drawn from a distribution. In the top row, this distribution is a mixture of two Gaussian distributions, one of unit variance at position [3,3] and the other of smaller variance (1/16) at position [-1,-1]. In the bottom row, training points are drawn from points along a sine wave. Anomaly detectors are derived in terms of the training data, using bandwidth $\sigma = 5$ (above: (a,b,c,d)) and $\sigma = 0.5$ (below: (e,f,g,h)). Anomalousness is scaled to range from zero to one, and plotted as white to black on a square that ranges from $[-B,B]$ on both axes. $B = 15$ above, and $B = 2$ below. Contours indicate false alarm rates of 0.05 (black), 0.01 (magenta), and 0.001 (cyan).

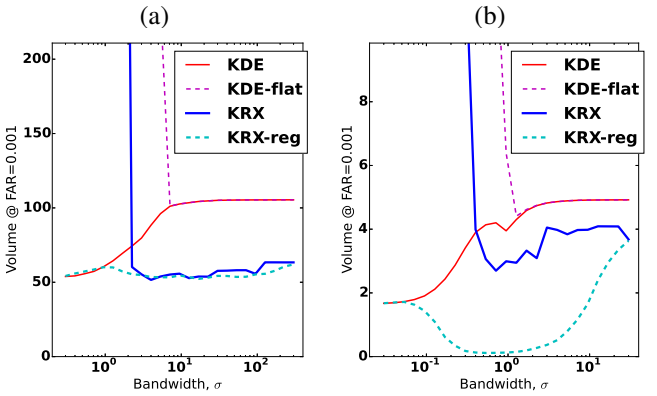

Fig. 3. Volume enclosed by anomaly detection contour associated with a false alarm rate of 0.001, plotted against bandwidth values. (a) Multi-modal Gaussian samples shown in the top row (a,b,c,d) of Fig. 2. (b) Sine wave samples shown in the bottom row (e,f,g,h) of Fig. 2. Smaller volumes are better.

KRX are unusable at small values of σ. As σ increases, the KDE contour becomes increasingly smooth and less able to effectively enclose the non-anomalous data. On the other hand, larger σ generally improves the performance of KRX.

This is also observed in Fig. 2(e,f,g,h) and Fig. 3(b), where KRX-reg strictly outperforms KDE-flat, KDE, and KRX over a wide range of bandwidth values σ.

C. Hyperspectral imagery

These projection effects are also evident in the application to hyperspectral imagery. In this subsection, we apply the kernelized algorithms (along with the non-kernelized RX)

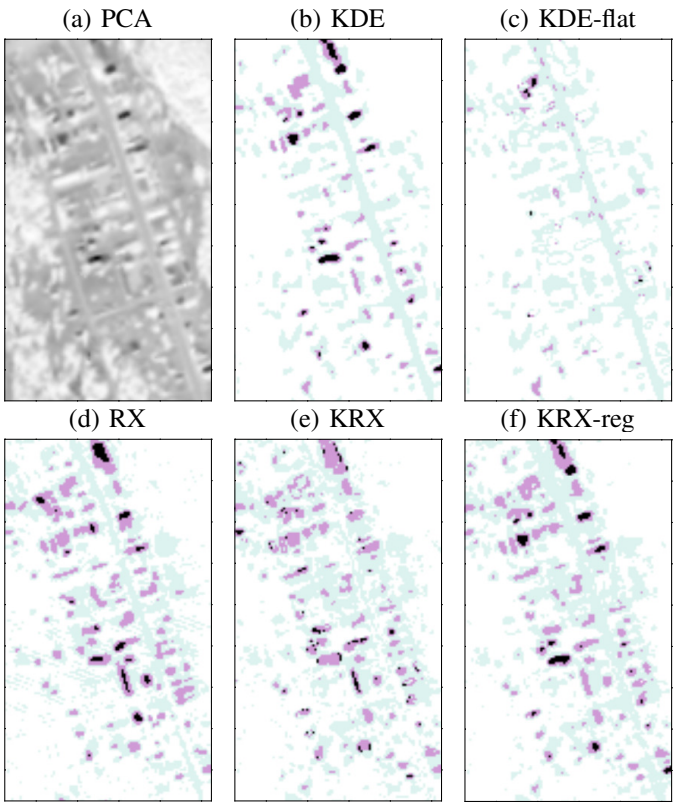

Fig. 4. The Cooke City HyMap hyperspectral dataset [14], is 280×800 pixels, with 126 spectral channels. The kernelized anomaly detectors were trained with $N = 1500$ randomly chosen pixels, using $\sigma = 5000$. (a) is the first principal component, (b-f) are the anomalousness maps for the different algorithms. The most anomalous 0.1% of the pixels are shown in black, the top 1% are pale magenta, and the top 10% are an even paler cyan. The analysis was applied to the full dataset, but shown here are 100×175 pixel insets.

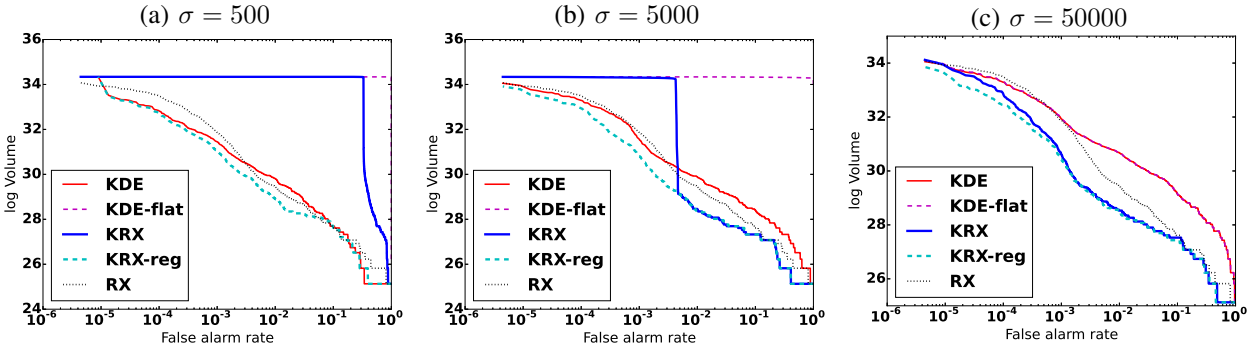

Fig. 5. For the Cooke City HyMap hyperspectral dataset, based on the first $D = 3$ principal components, we plot volume enclosed by the surface as a function of the false alarm rate (*i.e.,* the fraction of pixels in the image that would be declared anomalous at that threshold). This volume is estimated by uniformly filling a large ellipsoid (covering all the data and a generous margin beyond that as well) with 20000 random points, and computing the fraction of them whose anomalousness is within the given threshold.

to the widely used Cooke City dataset [14]. For numerical reasons, we employed a slight variant of the Gaussian kernel: $k'(\mathbf{r}, \mathbf{s}; \sigma) = k(\mathbf{r}, \mathbf{s}; \sigma) + \epsilon^2 k(\mathbf{r}, \mathbf{s}, \epsilon\sigma)$ with k defined in Eq. (3), and $\epsilon = 0.1$. This puts a little extra weight on the tails, but preserves the important properties of a kernel function [15].

Fig. 4 illustrates the differences that are observed when different anomaly detectors are applied to the same image. Comparing KRX to KRX-reg (and even more so, KDE-flat to KDE), we see that many of the most anomalous points lose their anomalousness in algorithms that employ a projection to the in-sample data subspace.

Of course, whether a given pixel is truly anomalous is a judgment call [16], so we also employed a more objective volume-based comparison of the different algorithms, as has been advocated previously [17], [18]. Because the volume of irregular contours is difficult to measure in high dimensions, we do the comparison using the first three principal components of the dataset. In Fig. 5, three values of sigma are used, and as was also observed in Fig. 3, we find that KDE prefers small σ, while KRX prefers larger σ. When σ is too small, the KRX volume diverges at low false alarm rate. But for adequately large σ, KRX outperforms RX and KDE. For all values of σ, however, the regularized KRX-reg is observed to perform well.

IV. Conclusion

Because of its projection to the in-sample subspace, KRX exhibits spuriously low anomalousness for points far from the training data. This same problem is observed in a simpler context (and more dramatic fashion) by comparing KDE-flat to KDE. Using ridge regularization (instead of pseudoinverse) on the covariance matrix in the feature space avoids this projection of the data. The result is KRX-reg, a kernelized anomaly detector that is as good or better than KRX and KDE over a wide range of bandwidth values.

Acknowledgments

JT was supported by the United States Department of Energy NA-22 Hyperspectral Advanced Research and Development for Solids project (HARD Solids). GG was supported by the Los Alamos Laboratory Directed Research and Development (LDRD) program.

References

[1] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, pp. 58–69, Jan 2002.

[2] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE A&E Systems Magazine*, vol. 25, pp. 5–27, 2010.

[3] P. C. Mahalanobis, "On the generalised distance in statistics," *Proc. National Institute of Sciences of India*, vol. 2, pp. 49–55, 1936.

[4] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1760–1770, 1990.

[5] D. Cremers, T. Kohlberger, and C. Schnörr, "Shape statistics in kernel space for variational image segmentation," *Pattern Recognition*, vol. 36, pp. 1929–1943, 2003.

[6] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: a nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, pp. 388–397, 2005.

[7] E. Parzen, "On estimation of probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

[8] B. W. Silverman, *Kernel Density Estimation Techniques for Statitistics and Data Analysis*. London: Chapman Hall, 1986.

[9] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *Ann. Statist.*, vol. 20, pp. 1236–1265, 1992.

[10] S. Matteoli, T. Veracini, M. Diani, and G. Corsini, "Background density nonparametric estimation with data-adaptive bandwidths for the detection of anomalies in multi-hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, pp. 163–167, 2014.

[11] ——, "A locally adaptive background density estimator: An evolution for RX-based anomaly detectors," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, pp. 323–327, 2014.

[12] N. M. Nasrabadi, "Kernel subspace-based anomaly detection for hyperspectral imagery," *1st IEEE Workshop on Hyperspectral Signal and Image Processing: Evolution in Remote Sensing (WHISPERS)*, 2009.

[13] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognition*, vol. 40, pp. 863–874, 2007.

[14] D. Snyder, J. Kerekes, I. Fairweather, R. Crabtree, J. Shive, and S. Hager, "Development of a web-based application to evaluate target finding algorithms," *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2, pp. 915–918, 2008.

[15] C. Scovel, D. Hush, I. Steinwart, and J. Theiler, "Radial kernels and their reproducing kernel Hilbert spaces," *Journal of Complexity*, vol. 26, pp. 641–660, 2010.

[16] J. Theiler, "By definition undefined: adventures in anomaly (and anomalous change) detection," *Proc. 6th IEEE Workshop on Hyperspectral Signal and Image Processing: Evolution in Remote Sensing (WHISPERS)*, 2014.

[17] D. Tax and R. Duin, "Uniform object generation for optimizing one-class classifiers," *J. Machine Learning Res.*, vol. 2, pp. 155–173, 2002.

[18] J. Theiler and D. Hush, "Statistics for characterizing data on the periphery," *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4764–4767, 2010.