CRACKS IN KRX: WHEN MORE DISTANT POINTS ARE LESS ANOMALOUS

James Theiler[*]

Los Alamos National Laboratory
Space Data Science and Systems Group
Los Alamos, NM 87545, USA

Guen Grosklos[†]

Utah State University
Mathematics & Statistics
Logan, UT 84322, USA

ABSTRACT

We examine the Mahalanobis-distance based kernel-RX (KRX) algorithm for anomaly detection, and find that it can exhibit an unfortunate phenomenon: the anomalousness, for points far from the training data, can decrease with increasing distance. We demonstrate this directly for a few special cases, and provide a more general argument that applies in the large bandwidth regime.

***Index Terms*—** Anomaly detection, Kernel density estimation, Mahalanobis distance, Kernel-RX

1. INTRODUCTION

For target detection generally, the key challenge is characterization of the background [1, 2]. In the anomaly detection problem, this challenge is expressed in a "pure" form that is isolated from the more domain-specific issues of target physics and target variability [3, 4].

Anomalies are data samples (*e.g.*, pixels in a hyperspectral image) that are unusual with respect to the rest of the data in a collection [5, 6, 7]. If the non-anomalous data samples are modeled by a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance R, then the Mahalanobis distance [8]

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

provides a simple measure of how anomalous the point $\mathbf{r}$ is. This is a measure that monotonically increases for increasing distance from the centroid $\boldsymbol{\mu}$.

Reed and Yu [9] used Mahalanobis distance to detect anomalies in multispectral and hyperspectral imagery, based on a local moving window, though the approach is often used globally as well, and in either case is popularly referred to as RX. Cremers *et al.* [10] proposed a kernelized version of RX, which Kwon and Nasrabadi [11] adopted for hyperspectral anomaly detection. Where RX provides elliptical contours of anomalousness, kernel-RX (KRX) can accommodate more convoluted contours.

In a recent paper [12], we identified an aspect of KRX – an implicit projection in feature space to the in-sample data plane – that leads in numerical studies to diminished performance, particularly at small bandwidths. In this paper, we explicitly show that KRX exhibits the peculiar property that for points far from the training data, anomalousness *decreases* with increasing distance.

2. KERNELIZATION

The training data has N samples, each d-dimensional: $\mathbf{x}_n \in \mathbb{R}^d$ for $n = 1, 2, \ldots, N$. For kernel-based methods, the data samples are mapped to a feature space $\mathcal{F}$ by a function Φ. That is $\Phi : \mathcal{X} \to \mathcal{F}$, where $\mathcal{F}$ has the property that dot products can be expressed as a scalar function of points in the original data space:

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

The "kernel trick" recognizes that by specifying the kernel function $\kappa(\mathbf{r}, \mathbf{s})$, one need not actually evaluate $\Phi(\mathbf{r})$ or $\Phi(\mathbf{s})$. We consider in particular the Gaussian radial basis kernel:

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

where σ is called the bandwidth.

Define $\boldsymbol{\mu}_\Phi$ as the centroid of the sample data in feature space: $\boldsymbol{\mu}_\Phi = \frac{1}{N} \sum_{n=1}^{N} \Phi(\mathbf{x}_n)$. In terms of this centroid, 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:

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

Note that this equation was written incorrectly in Ref. [12].

[*]JT was supported by the United States Department of Energy (DOE) NA-22 Hyperspectral Advanced Research and Development for Solids (HARD Solids) project.

[†]GG performed the work while at Los Alamos National Laboratory, supported by the Los Alamos Laboratory Directed Research and Development (LDRD) program.

2.1. Kernel-RX (KRX)

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

$$X_\Phi = [\Phi_c(\mathbf{x}_1) \cdots \Phi_c(\mathbf{x}_N)], \tag{5}$$

Let r be the rank of this matrix (observe that $r \leq N-1$ since the centroid has been subtracted). Express X_Φ with a singular value decomposition

$$X_\Phi = V_\Phi \Lambda^{1/2} W^T. \tag{6}$$

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 $C_\Phi = X_\Phi X_\Phi^T$, and columns of W are eigenvectors of the centered Gram matrix

$$K_c = X_\Phi^T X_\Phi = \begin{bmatrix} \kappa_c(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \kappa_c(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \kappa_c(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \kappa_c(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \tag{7}$$

Note that this $N \times N$ matrix has rank at most $N-1$ (since that is the rank of X_Φ); thus it is not strictly invertible.

The KRX idea is to use a Mahalanobis distance in the feature space as the measure of anomalousness. That is:

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \Phi_c(\mathbf{r})^T C_\Phi^{-1} \Phi_c(\mathbf{r}), \tag{8}$$

with the covariance matrix C_Φ given by

$$C_\Phi = \sum_n \Phi_c(\mathbf{r}) \Phi_c(\mathbf{r})^T = X_\Phi X_\Phi^T = V_\Phi \Lambda V_\Phi^T \tag{9}$$

where X_Φ was defined in Eq. (5), and decomposed in Eq. (6). The problem with KRX, as it is expressed in Eq. (8), is that C_Φ is not invertible. The approach taken in [11] is to use the pseudoinverse. That is,

$$(V_\Phi \Lambda V_\Phi^T)^{-1} \leftarrow V_\Phi \Lambda^{-1} V_\Phi^T. \tag{10}$$

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. (6), to further simplify

$$\begin{aligned} C_\Phi^{-1} &= (X_\Phi W \Lambda^{-1/2}) \Lambda^{-1} (\Lambda^{-1/2} W^T X_\Phi^T) \\ &= X_\Phi W \Lambda^{-2} W^T X_\Phi^T = X_\Phi K_c^{-2} X_\Phi^T, \end{aligned} \tag{11}$$

where K_c is the centered Gram matrix defined in Eq. (7), and K_c^{-2} refers to the pseudoinverse of K_c^2. Thus,

$$\begin{aligned} \mathcal{A}_{\text{KRX}}(\mathbf{r}) &= \Phi_c(\mathbf{r})^T X_\Phi K_c^{-2} X_\Phi^T \Phi_c(\mathbf{r}) \\ &= \mathbf{z}_c(\mathbf{r})^T K_c^{-2} \mathbf{z}_c(\mathbf{r}), \end{aligned} \tag{12}$$

where $\mathbf{z}_c(\mathbf{r})$ can be expressed in terms of the centered kernel that was derived in Eq. (4):

$$\mathbf{z}_c(\mathbf{r}) = X_\Phi^T \Phi_c(\mathbf{r}) = \begin{bmatrix} \kappa_c(\mathbf{x}_1, \mathbf{r}) \\ \vdots \\ \kappa_c(\mathbf{x}_N, \mathbf{r}) \end{bmatrix}. \tag{13}$$

If we define the scalar $g(\mathbf{r}) = \frac{1}{N} \sum_n \kappa(\mathbf{x}_n, \mathbf{r})$, and the vector

$$\mathbf{k}(\mathbf{r}) = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{r}) - g(\mathbf{r}) \\ \vdots \\ \kappa(\mathbf{x}_N, \mathbf{r}) - g(\mathbf{r}) \end{bmatrix}, \tag{14}$$

and further define

$$\mathbf{k}_\mu = \frac{1}{N} \sum_n \mathbf{k}(\mathbf{x}_n), \tag{15}$$

then $\mathbf{z}_c(\mathbf{r}) = \mathbf{k}(\mathbf{r}) - \mathbf{k}_\mu$, and the anomalousness in Eq. (12) can be expressed as

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \left(\mathbf{k}(\mathbf{r}) - \mathbf{k}_\mu\right)^T K_c^{-2} \left(\mathbf{k}(\mathbf{r}) - \mathbf{k}_\mu\right), \tag{16}$$

which corresponds to the Mahalanobis distance in an empirical kernel space defined by the map $\mathbf{r} \to \mathbf{k}(\mathbf{r})$.

3. SIMPLE CASE: TWO TRAINING SAMPLES

As an initial example, we consider the special case of a training data set with two points in $\mathbb{R}^d$. This is the simplest informative example. In addition to providing a specific scenario that can be explicitly analyzed, it will also shed some light on what the phenomenon looks like in the more general (and complicated) case.

Without loss of generality, we can rotate, translate, and scale the data so that the two training points are $\mathbf{x}_- = (-1, 0)$ and $\mathbf{x}_+ = (+1, 0)$, and consider only the space $\mathbb{R}^2$ containing the points $\mathbf{x}_-$ and $\mathbf{x}_+$ and the point $\mathbf{r} = (x, y)$ at which anomalousness is measured.

3.1. Centered Gram matrix

The kernel $\kappa(\mathbf{x}, \mathbf{x}) = 1$ when the two arguments are equal, and $\kappa(\mathbf{x}_+, \mathbf{x}_-) = \kappa(\mathbf{x}_-, \mathbf{x}_+) = \exp(-2/\sigma^2)$, and we will write $\theta \equiv \exp(-2/\sigma^2)$ for convenience. Then the *centered* kernel defined in Eq. (4) is given by:

$$\kappa_c(\mathbf{x}_-, \mathbf{x}_+) = \kappa_c(\mathbf{x}_+, \mathbf{x}_-) = -(1-\theta)/2 \tag{17}$$

$$\kappa_c(\mathbf{x}_-, \mathbf{x}_-) = \kappa_c(\mathbf{x}_+, \mathbf{x}_+) = (1-\theta)/2 \tag{18}$$

Hence

$$K_c = \frac{1}{2}(1-\theta) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{19}$$

Since K_c is not invertible, the KRX formula in Eq. (12) cannot be applied directly. As derived in Section 2.1, the pseudoinverse of K_c^2 is the appropriate choice for K_c^{-2}, but we will also consider the inverse of a ridge regularized K_c^2.

It is straightforward to show that the pseudoinverse is

$$K_c^{-2} \leftarrow \text{pinv}(K_c^2) = \frac{1}{2(1-\theta)^2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad (20)$$

and that the ridge-regularized inverse is given by

$$K_c^{-2} \leftarrow (K_c^2 + \lambda I)^{-1} = \frac{1}{2[\lambda + (1-\theta)^2]}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{1}{2\lambda}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad (21)$$

3.2. Anomalousness at r

For a general point $\mathbf{r}$, we can write the scalar

$$g(\mathbf{r}) = \frac{1}{2}\left[\kappa(\mathbf{x}_-, \mathbf{r}) + \kappa(\mathbf{x}_+, \mathbf{r})\right] \quad (22)$$

so that the vector $\mathbf{k}(\mathbf{r})$ defined in Eq. (14) becomes

$$\mathbf{k}(\mathbf{r}) = \begin{bmatrix} \kappa(\mathbf{x}_-, \mathbf{r}) - g(\mathbf{r}) \\ \kappa(\mathbf{x}_+, \mathbf{r}) - g(\mathbf{r}) \end{bmatrix} = \zeta(\mathbf{r})\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad (23)$$

where

$$\zeta(\mathbf{r}) = \frac{1}{2}\left[\kappa(\mathbf{x}_-, \mathbf{r}) - \kappa(\mathbf{x}_+, \mathbf{r})\right]. \quad (24)$$

Since $\zeta(\mathbf{x}_-) + \zeta(\mathbf{x}_+) = 0$, it is clear that $\mathbf{k}_\mu = 0$. Finally, using the expression for $\mathbf{k}(\mathbf{r})$ in Eq. (23), substituted into Eq. (16), we obtain

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \zeta^2(\mathbf{r})\begin{bmatrix} 1 & -1 \end{bmatrix} \times K_c^{-2} \times \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{2\zeta^2(\mathbf{r})}{\lambda + (1-\theta)^2}, \quad (25)$$

where we have used Eq. (21) for K_c^{-2} and the identities:

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \times \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 4, \quad \text{and} \quad (26)$$

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0. \quad (27)$$

Observe that the $\lambda \to 0$ limit of Eq. (25) is what would be obtained if the pseudoinverse were used for K_c^{-2} instead of the ridge-regularized inverse.

For a position $\mathbf{r} = (x, y)$, we can write

$$\kappa(\mathbf{x}_\pm, \mathbf{r}) = \exp\left(-[(x \mp 1)^2 + y^2]/2\sigma^2\right); \quad (28)$$

thus Eq. (24) becomes

$$\zeta(x, y) = \frac{1}{2}\exp\left(-[1 + x^2 + y^2]/2\sigma^2\right) \times \left[\exp(-2x/2\sigma^2) - \exp(2x/2\sigma^2)\right] \quad (29)$$

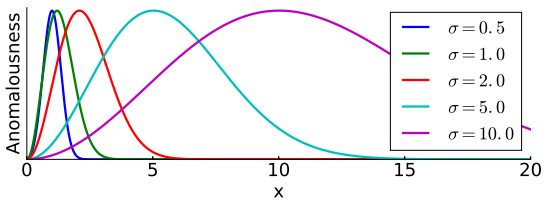

Fig. 1. Plot of $\mathcal{A}_{\text{KRX}}(x, y)$, as defined in Eq. (30), against x for bandwidth values $\sigma = 0.5, 1, 2, 5, 10$. Anomalousness initially increases until approximately $x = \sigma + 1/6\sigma$, after which it decreases, approaching zero as $x \to \infty$.

and so

$$\mathcal{A}_{\text{KRX}}(x, y) \propto \exp(-(x^2 + y^2)/\sigma^2)\sinh^2(x/\sigma^2). \quad (30)$$

This expression is non-negative (and strictly positive except at $x = 0$), and approaches zero as $|x|$ or $|y|$ go to infinity. Thus, for large $|x|$, we have that $\mathcal{A}_{\text{KRX}}(x, y)$ decreases with increasing distance from the origin. In fact, as Fig. 1 shows, $\mathcal{A}_{\text{KRX}}(x, y)$ initially increases with increasing x, providing a reasonable measure of anomalousness as long as $x < \sigma$. But after that, anomalousness decreases with increasing x.

4. SIMPLE CASE: N POINTS ON A SIMPLEX

Another simple case considers N points in $\mathbb{R}^d$ for $d \geq N-1$, that are points on the corners of a regular simplex. For these points, all pairwise distances are equal, and we will again write $\theta = \exp(-\|x_i - \mathbf{x}_j\|^2/2\sigma^2)$ for $i \neq j$. Note that $N = 2$ reverts to the example in the previous section. We can compute:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \theta + (1-\theta)\delta_{ij}$$
$$g(\mathbf{x}_i) = \frac{1}{N}\sum_n \kappa(\mathbf{x}_n, \mathbf{x}_i) = \theta + (1-\theta)/N$$
$$k_i(\mathbf{x}_j) = \kappa(\mathbf{x}_i, \mathbf{x}_j) - g(\mathbf{x}_j) = (1-\theta)\left[\delta_{ij} - 1/N\right].$$

where $k_i(\mathbf{r})$ is the ith component of the vector $\mathbf{k}(\mathbf{r})$ defined in Eq. (14). It follows that the ith component of the vector $\mathbf{k}_\mu$ is $k_{\mu i} = \frac{1}{N}\sum_n k_i(\mathbf{x}_n) = 0$. With $\mathbf{k}_\mu = 0$, Eq. (16) becomes $\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \mathbf{k}(\mathbf{r})^T K_c^{-2}\mathbf{k}(\mathbf{r})$, which is everywhere non-negative and approaches zero as $\mathbf{r} \to \infty$. Thus, more distant points $\mathbf{r}$ are less anomalous. Fig. 2 shows anomalousness initially increasing with distance away from the training samples, but then decreasing for larger $\mathbf{r}$.

5. LARGE σ REGIME

In this section, we consider a regime in which σ is much larger than the distances between the training points. Although our approach will be less formal than in previous sections, it will also make fewer demands on the layout of the training samples. What we will find is that $\mathcal{A}_{\text{KRX}}(\mathbf{r})$ generically takes on

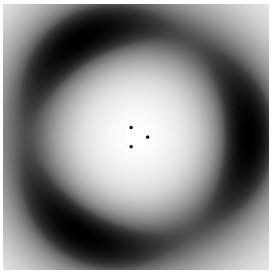

Fig. 2. Anomalousness in $\mathbb{R}^2$ for a simplex of $N = 3$ points (indicated as black dots) on the unit circle. Darker shades indicate higher anomalousness. The bandwidth is $\sigma = 10$, much larger than the distance between points, and the range shown is $[-12, 12]$.

values that tend to be much larger when $\|\mathbf{r}\| = O(\sigma)$ than the asymptotic value as $\mathbf{r} \to \infty$.

Without loss of generality, we will translate and scale the training points so that $\sum_n \mathbf{x}_n = 0$ and $\frac{1}{N}\sum_n \|\mathbf{x}_n\|^2 = 1$.

5.1. Magnitude of $\mathbf{k}_\mu$

Since $\|\mathbf{x}_i - \mathbf{x}_j\| \ll \sigma$, we can write

$$\begin{aligned}\kappa(\mathbf{x}_i, \mathbf{x}_j) &= \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2) \\ &= 1 - \|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2 + O(1/\sigma^4) \\ &\approx 1 - \|\mathbf{x}_i\|^2/2\sigma^2 - \|\mathbf{x}_j\|^2/2\sigma^2 - \mathbf{x}_i \cdot \mathbf{x}_j/\sigma^2\end{aligned}$$

using the identity $\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i \cdot \mathbf{x}_j$. We then use both $\sum_n \mathbf{x}_n = 0$ and $\sum_n \|\mathbf{x}_n\|^2 = 1$ to write

$$g(\mathbf{x}_i) = \frac{1}{N}\sum_n \kappa(\mathbf{x}_i, \mathbf{x}_n) \approx 1 - 1/2\sigma^2 - \|\mathbf{x}_i\|^2/2\sigma^2.$$

Let us write $k_j(\mathbf{x}_i)$ as the jth component of $\mathbf{k}(\mathbf{r})$ for $\mathbf{r} = \mathbf{x}_i$:

$$\begin{aligned}k_j(\mathbf{x}_i) &= \kappa(\mathbf{x}_i, \mathbf{x}_j) - g(\mathbf{x}_i) \\ &\approx (1 - \|\mathbf{x}_j\|^2)/2\sigma^2 + 2\mathbf{x}_i \cdot \mathbf{x}_j/2\sigma^2\end{aligned}$$

And finally,

$$k_{\mu j} = \frac{1}{N}\sum_n k_j(\mathbf{x}_n) \approx (1 - \|\mathbf{x}_j\|^2)/2\sigma^2.$$

Thus, $\mathbf{k}_\mu = O(1/\sigma^2)$.

5.2. Magnitude of $\mathbf{k}(\mathbf{r})$ for $\|\mathbf{r}\| \approx \sigma$

Use $\|r - \mathbf{x}\|^2 = \|\mathbf{r}\|^2 + \|\mathbf{x}\|^2 - 2\mathbf{r} \cdot \mathbf{x}$ to write

$$\begin{aligned}\kappa(\mathbf{x}_i, \mathbf{r}) &= \exp(-\|\mathbf{r} - \mathbf{x}_i\|^2/2\sigma^2) \\ &= \exp(-\|\mathbf{r}\|^2/2\sigma^2) \times \exp(-\|\mathbf{x}_i\|^2/2\sigma^2) \\ &\quad \times \exp(-2\mathbf{r} \cdot \mathbf{x}_i/2\sigma^2) \\ &= \exp(-\|\mathbf{r}\|^2/2\sigma^2)[1 - 2\mathbf{r} \cdot \mathbf{x}_i/2\sigma^2] + O(1/\sigma^2)\end{aligned}$$

So then

$$g(\mathbf{r}) = \frac{1}{N}\sum_n \kappa(\mathbf{x}_n, \mathbf{r}) \approx \exp(-\|\mathbf{r}\|^2/2\sigma^2)$$

and

$$k_j(\mathbf{r}) = \kappa(\mathbf{x}_j, \mathbf{r}) - g(\mathbf{r}) \approx \underbrace{\exp(-\|\mathbf{r}\|^2/2\sigma^2)}_{O(1)}\underbrace{[-2\mathbf{r} \cdot \mathbf{x}_j/2\sigma^2]}_{O(1/\sigma)}.$$

Thus, for $\|\mathbf{r}\| \approx \sigma$, we have $\mathbf{k}(\mathbf{r}) = O(1/\sigma)$.

5.3. Anomalousness for $\|\mathbf{r}\| \approx \sigma$ and for $\mathbf{r} \to \infty$

From Eq. (16), we can write

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) = \|K_c^{-1}\left[\mathbf{k}(\mathbf{r}) - \mathbf{k}_\mu\right]\|^2 \tag{31}$$

$$\sim \|K_c^{-1}\|^2 \times \|\mathbf{k}(\mathbf{r}) - \mathbf{k}_\mu\|^2 \tag{32}$$

where the second line is not a formal statement, but expresses the general magnitudes of the different terms. In particular,

$$\mathcal{A}_\infty \sim \|K_c^{-1}\|^2 \times \|\mathbf{k}_\mu\|^2 = \|K_c^{-1}\|^2 \times O(1/\sigma^4).$$

By contrast, for $\|\mathbf{r}\| = O(\sigma)$, we have $\mathbf{k}(r) = O(1/\sigma)$, which in general dominates $\mathbf{k}_\mu = O(1/\sigma^2)$. In this regime,

$$\mathcal{A}_{\text{KRX}}(\mathbf{r}) \sim \|K_c^{-1}\|^2 \times \|\mathbf{k}(\mathbf{r})\|^2 = \|K_c^{-1}\|^2 \times O(1/\sigma^2).$$

By a factor of $\sigma^2 \gg 1$, we see that $\mathcal{A}_{\text{KRX}}(\mathbf{r})$ values tend to be much larger for $\mathbf{r} \sim \sigma$ than for $\mathbf{r} \to \infty$. While KRX might behave reasonably for $\|\mathbf{r}\| < \sigma$, this result shows that for larger $\mathbf{r}$, the anomalousness $\mathcal{A}_{\text{KRX}}(\mathbf{r})$ tends to *decrease* as the distance from $\mathbf{r}$ to the training data *increases*.

6. CONCLUSIONS

We have demonstrated in very specific cases that the kernel-RX algorithm exhibits an unfortunate property for anomaly detectors: points that are farther from the normal data are less anomalous. We have not shown this for arbitrary training data, but we speculate – based both on these special cases and on our (albeit anecdotal) experience with numerical computation – that this property is ubiquitous.

This property is not shared by all kernel-based anomaly detectors: kernel density estimation [13, 14], support vector data decomposition [15, 16] and kernel principal components analysis [17, 18] behave more rationally in this respect, and identify points farther from the training data as more anomalous. This work extends a previous result [12], based on numerical evidence, which traced the source of this problem to an implicit projection to the in-sample subspace that occurs in Eq. (10) when the pseudoinverse of C_Φ is taken. Modifying KRX by regularizing the C_Φ matrix before inverting it (note that regularizing K_c^2 is not enough), leads to a kernelized anomaly detector that avoids this unfortunate property [10, 12].

In practice, this is not necessarily a showstopper. If the bandwidth σ is larger than the largest distance $\mathbf{r}$ that will be considered, then KRX can provide reasonable contours of anomalousness. Indeed, the $\sigma \to \infty$ limit for KRX yields straight RX. But however large σ is, there will be points $\mathbf{r}$ for which anomalousness decreases with increasing distance.

7. REFERENCES

[1] S. Matteoli, M. Diani, and J. Theiler, "An overview background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery," *IEEE J. Sel. Topics in Applied Earth Observations and Remote Sensing*, vol. 7, pp. 2317–2336, 2014.

[2] J. Theiler and B. Wohlberg, "Regression framework for background estimation in remote sensing imagery," *5th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2013.

[3] T. L. Myers, C. S. Brauer, Y.-F. Su, T. A. Blake, R. G. Tonkyn, A. B. Ertel, T. J. Johnson, and R. L. Richardson, "Quantitative reflectance spectra of solid powders as a function of particle size," *Applied Optics*, vol. 54, pp. 4863–4875, 2015.

[4] A. K. Ziemann, J. Theiler, and D. W. Messinger, "Hyperspectral target detection using manifold learning and multiple target spectra," *Proc. 44th IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2015.

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] J. Theiler and G. Grosklos, "Problematic projection to the in-sample subspace for a kernelized anomaly detector," *IEEE Geoscience and Remote Sensing Lett.*, vol. 13, pp. 485–489, 2016.

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

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

[15] D. Tax and R. Duin, "Data domain description by support vectors," in *Proc. ESANN99*, M. Verleysen, Ed. Brussels: D. Facto Press, 1999, pp. 251–256.

[16] A. Banerjee, P. Burlina, and C.Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Trans. Geoscience and Remote Sensing*, vol. 44, pp. 2282–2291, 2006.

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

[18] 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.