Anomaly testing

James Theiler
Space Data Science and Systems Group
Los Alamos National Laboratory

Abstract

An anomaly is something that in some unspecified way stands out from its background, and the goal of anomaly testing is to determine which samples in a population stand out the most. This chapter presents a conceptual discussion of anomalousness, along with a survey of anomaly detection algorithms and approaches, primarily in the context of hyperspectral imaging. In this context, the problem can be roughly described as target detection with unknown targets. Because the targets, however tangible they may be, are unknown, the main technical challenge in anomaly testing to characterize the background. Further, because anomalies are rare, it is the characterization not of a full probability distribution but of its periphery that most matters. One seeks not a generative but a discriminative model, an envelope of unremarkability, the outer limits of what is normal. Beyond those outer limits, *hic sunt anomalias.*

CONTENTS

Appears as: Chapter 19 in *Statistical Methods for Materials Science: The Data Science of Microstructure Characterization*, J. P. Simmons, C. A. Bouman, M. De Graef, and L. F. Drummy, Jr., eds. (CRC Press, 2018). ISBN 9781498738200.

I. Introduction

Traditionally, anomalies are defined in a negative way, not by what they are but by what they are not: they are data samples that are not like the data samples in the rest of data. "There is not an unambiguous way to way to define an anomaly," one review notes, and then goes on to ambiguously define it as "an observation that deviates in some way from the background clutter" [1]. Anomalies are defined "without reference to target signatures or target subspaces" and "with reference to a model of the background" [2]. Indeed, as Ashton [3] remarks, "the basis of an anomaly detection system is accurate background characterization."

This exposition will concentrate on anomaly detection in the context of imagery, with particular emphasis on hyperspectral imagery (in which each pixel encodes not the usual red, green, and blue of visible images, but a spectrum of radiances over a range of wavelengths that often includes upwards of a hundred spectral channels). Testing for anomalies is an exercise that has application in a variety of scenarios, however. Since anomalies are deviations from what is normal, particularly in situations where the nature of that deviation is not be predictable or well characterized, anomaly detection has been used in a variety of fault detection contexts [4]–[8].

What we are calling anomaly testing here is essentially the same as what the machine learning community calls "novelty detection" [9]–[11] or "one-class classification" [12], [13].

A. Anomaly testing as triage

Although we may have difficulty *defining* anomalies, the reason we seek them is that (being rare and unlike most of the data) they are potentially interesting and possibly meaningful. We have to acknowledge that "interesting" is even harder to define than "anomalous" but in this exposition, we will make this distinction, partly because it separates the mystical "by definition undefined" [14] aspect of the interesting-data detection problem into two components, which correspond to the two boxes in Fig. 1.

We can indeed define "anomalous," but will leave "interesting" and "meaningful" to be domain-specific concepts. From this point of view, anomaly detection is a kind of triage. If anomalies can be defined and detected in a relatively generic way, then experts in the specific area of application can decide which anomalies are interesting or meaningful.

Here the goal of anomaly detection is to reduce the quantity of incoming data to a level that can be handled by the more expensive downstream analysis. It is this later analysis that judges which of the anomalous items are in fact meaningful for the application at hand. This judgment can be very complicated and domain-specific, and can involve human acumen and intuition. What makes anomaly detection useful as a concept is that the anomaly detection module has more generic goals, and is consequently more amenable to formal mathematical analysis.

B. Anomalies drawn from a uniform distribution

Anomalies are rare, and where we expect to find them is in the far tails of the background distribution $p_b(\mathbf{x})$. We can express "anomalousness" as varying inversely with with this density function, and can derive this expression in two distinct ways, each providing its own insight.

In the first and most direct approach, we make an explicit generative model for anomalies, and say that they are samples drawn from a uniform distribution. This is a simple statement, but it is in some ways revolutionary; in contrast to conventional wisdom [1]–[3], we are defining anomalies directly, *without respect* to the background distribution. To distinguish these anomalies from the background, we treat the detection as a hypothesis testing problem. The null hypothesis

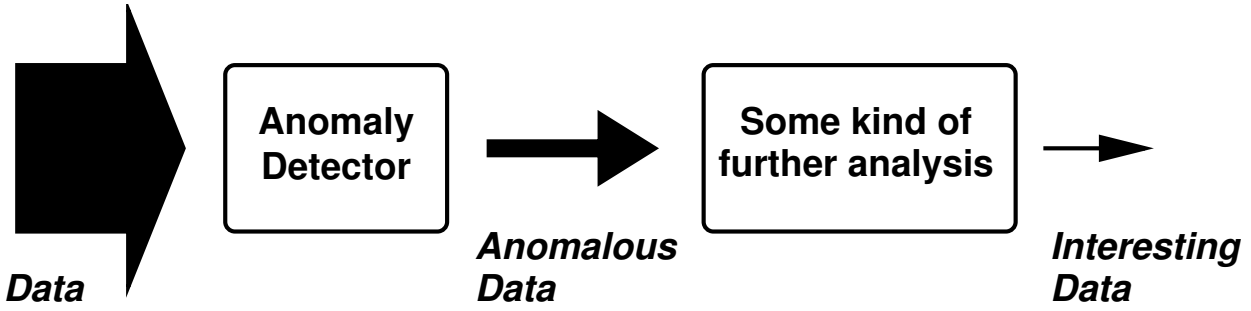


Fig. 1. Anomaly testing as triage. In this picture, anomaly detection provides a way to reduce the raw quantity of data that needs to be considered by further analysis. That further analysis might be expensive domain-specific automated processing, or it may involve human inspection, with trained analysts making judgments about whether the *potentially* interesting anomaly is indeed interesting or meaningful or important.

is that the measurement $\mathbf{x}$ is drawn from the background distribution ($\mathbf{x} = \mathbf{z}$ with $\mathbf{z} \sim p_b$), and the alternative hypothesis is that $\mathbf{x}$ is drawn from the anomalous distribution ($\mathbf{x} = \mathbf{t}$ with $\mathbf{t} \sim u$). This leads to the likelihood ratio

$$\mathcal{L}(\mathbf{x}) = \frac{P(\mathbf{x} = \mathbf{t})}{P(\mathbf{x} = \mathbf{z})} = \frac{u(\mathbf{x})}{p_b(\mathbf{x})} = \frac{c}{p_b(\mathbf{x})} \tag{1}$$

where $u(\mathbf{x}) = c$ is the uniform distribution from which anomalies are drawn.[1] The expression in Eq. (1) is a candidate for "anomalousness" because a large value of $\mathcal{L}(\mathbf{x})$ indicates a higher likelihood that $\mathbf{x}$ is drawn from the anomalous distribution.

In the second approach, we avoid making an explicit model of what an anomaly is, but we assume that the effect of the anomaly on the scene is additive. In particular, we say that a pixel with an anomaly in it is of the form $\mathbf{x} = \mathbf{z} + \mathbf{t}$ where $\mathbf{t}$ is the unknown target, and $\mathbf{z}$ is the background. We employ the *generalized* likelihood ratio test to write

$$\mathcal{L}(\mathbf{x}) = \frac{P(\mathbf{x} = \mathbf{z} + \mathbf{t})}{P(\mathbf{x} = \mathbf{z})} = \frac{\max_{\mathbf{t}} p_b(\mathbf{x} - \mathbf{t})}{p_b(\mathbf{x})} = \frac{\max_{\mathbf{z}} p_b(\mathbf{z})}{p_b(\mathbf{x})} = \frac{c'}{p_b(\mathbf{x})} \tag{2}$$

where the (again, irrelevant) constant c' is the maximum value of p_b and does not depend on $\mathbf{x}$. This second approach produces the same result as the first, but the argument used to get that result has two problems: one, the additive assumption is very restrictive and may not apply to the scenario of interest; and two, it uses a generalized likelihood ratio test (GLRT). Although the GLRT is a popular and often practical tool, it has ambiguous properties, and can produce detectors that are not only sub-optimal, but in some cases *inadmissible* [16], [17]. This is not to say that the detector in Eq. (2) is inadmissible; indeed, it is identical to the detector in Eq. (1). The objection is not to the detector but to this second derivation of the detector *via* the GLRT. A supposed advantage of this second derivation is that it makes "no assumptions" about the nature of the anomaly, but this is a spurious argument. In fact, it is making implicit assumptions about the distribution of an anomaly, but it is not providing the algorithm designer with access to alter those implicit assumptions.

Although there are many reasons to favor the first approach, it is the second argument that is most widely invoked in the hyperspectral anomaly detection literature.

[1]The constant c is irrelevant to our purposes, and in fact it is possible to make this argument in a more formal way that treats $u(\mathbf{x})$ not as a proper probability distribution, but more generally as a measure [15].

We remark that the first approach can also accommodate additive anomalies. Here, the numerator of the likelihood ratio is a Bayes factor, but it still evaluates to a constant independent of $\mathbf{x}$:

$$\mathcal{L}(\mathbf{x}) = \frac{P(\mathbf{x} = \mathbf{z} + \mathbf{t})}{P(\mathbf{x} = \mathbf{z})} = \frac{\int p_b(\mathbf{x} - \mathbf{t})u(\mathbf{t})d\mathbf{t}}{p_b(\mathbf{x})} = \frac{c\int p_b(\mathbf{z})d\mathbf{z}}{p_b(\mathbf{x})} = \frac{c''}{p_b(\mathbf{x})} \tag{3}$$

We again obtain the result that anomalousness varies inversely with $p_b(\mathbf{x})$, the probability density function of the background. Contours of anomalousness will be level curves of the background density functions. For a Gaussian distribution, these contours are ellipsoids of constant Mahalanobis distance [18], with larger distances corresponding to smaller densities and greater anomalousness; we can therefore use Mahalanobis distance as a measure of anomalousness

$$\mathcal{A}(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T R^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4}$$

where $\boldsymbol{\mu}$ is a vector-valued mean and R is a covariance matrix.

The Mahalanobis distance is the basis of the Reed-Xiaoli (RX) detector [19]–[21]. Although RX, as originally introduced [19], refers specifically to multispectral imagery, and in fact is a local anomaly detector, the term "RX" is often used as a shorthand for Mahalanobis distance based anomaly detection.

1) Nonuniform distributions of anomalousness: Since the assumption of a uniform distribution for anomalies is explicit in the derivation of Eq. (1) – and by extension, Eq. (4) – that allows us to consider other, nonuniform, distributions if we are looking for other kinds of anomalies. Examples include anomalous change (to be further discussed in Section VI), and anomalous "color" in multispectral imagery. Here, in place of a uniform distribution of anomalies, a distribution of anomalously colored pixels are generated from the product of marginal distributions associated with each individual spectral band. To sample at random from this distribution, one creates a vector-valued pixel where each component is independently sampled from the corresponding component of the multispectral image [22].

Another example is given by the blind gas detection algorithm. In the traditional gas detection problem, one is looking for plumes that contain a gas-phase chemical of interest. When the absorption spectrum of that chemical of interest is known, one can derive a matched filter that can detect impressively low concentrations of the chemical [23]. In the blind gas detection problem, one does not have a single chemical (or even a short list of chemicals) of interest; one wants to detect chemical plumes without knowing the chemical species in the plume. But one does know that gas-phase chemicals usually have very sparse spectral signatures – *i.e.*, there is zero or nearly-zero absorption at all but a few wavelengths. A sparse RX (or "spaRX") algorithm was developed for detecting spectrally sparse additive anomalies $\mathbf{t}$ based on the RX derived in Eq. (2), but with $\mathbf{t}$ constrained to a limited number of nonzero components [24].

C. Anomalies as pixels in spectral imagery

Traditional statistical analysis treats data as a set of discrete samples that are drawn from a common distribution. Because each pixel in a hyperspectral image contains so much information (a many-channel spectrum of reflectances or radiances), one can often quite profitably treat the pixels as independent and identically distributed. It is as if the image were a "bag of pixels." But however spectrally informative individual pixels are, they comprise an image, and the spatial structure in an image provides further leverage for characterizing the background and discovering anomalies.

Hyperspectral imagery provides a rich and irregular data set, with complex spatial and spectral structure. And the more accurately we can model this cluttered background, the better our detection performance. Simple models can be very effective, but the mismatch between simple models and the complicated nature of real data has driven research toward the development of more complex models [25].

1) Global and local anomaly detectors: A pixel is anomalous in the context of a background. In *global* anomaly detection that background is the full image, but in *local* anomaly detection that background restricted to the immediate neighborhood, often defined in terms of an annulus that surrounds the pixel. Local anomaly detection is one of the most straightforward (and, in practice, more effective) ways to exploit the spatial structure of imagery.

Indeed, the initial formulation of the RX algorithm [19] computed anomalousness at each pixel in terms of a local mean and a local covariance matrix, each computed from the pixels in an annulus surrounding that pixel. The trade-off in choosing the size of the annulus is that a larger annulus will have better "statistics" – since it has more pixels, it will better average out the fluctuations in the pixels values; but a smaller annulus will be less affected by spatial nonstationarity [26]. Matteoli *et al.* [1] observed that one could use a smaller annulus for the local mean and a larger annulus for the local covariance. This is very sensible, since the need for "good statistics" is greater for the covariance matrix than for the mean vector. As the covariance annulus approaches the size of the image, this approach can be simplified by using a local mean and a global covariance, but with the global covariance based on subtraction of the spatially varying local mean. A broad survey of approaches used to improve estimates of local covariance, including regularization, segmentation, and robustification, is provided by Matteoli *et al.* [27].

The importance of regularization derives from the fact that RX requires the *inverse* of the covariance matrix. If the covariance matrix is singular, then the inverse does not exist, and regularization is required. But even for a well-conditioned covariance matrix, the best estimator of the inverse is not the same as the inverse of the best estimator, and some amount of regularization is still beneficial. The most common and straightforward regularization is by shrinkage. Here, we estimate the covariance matrix with a linear combination of the sample covariance R_s (which tends to overfit the data) and a very simple estimator R_o that tends to underfit the data

$$\widehat{R} = (1-\alpha)R_s + \alpha R_o \tag{5}$$

where typically $\alpha \ll 1$. In the simplest case, R_o is just a multiple of the identify matrix [28], [29] (choosing the multiple so that R_o has trace equal to R_s ensures that α is dimensionless). An argument can be made for shrinkage against the diagonal matrix [30], [31], an approach that is generalized in the sparse matrix transform [32], [33]. Caefer *et al.* [34] recommended a quasi-local estimator that combines local and global covariance estimators by using local eigenvalues with global eigenvectors.

The idea of segmentation is to replace the moving window with a static segment of similar pixels that surround the pixel of interest in a more irregular way. Here the image is partitioned into distinct segments of (usually contiguous) pixels, and a pixel's anomalousness is based on the mean and covariance of the pixels in the segment to which the pixel belongs [34], [35]. This sometimes leads to extra false alarms on the boundaries of the segments, and one way to deal with this problem is with overlapping segments [36].

The estimation of covariance in the local annulus can be corrupted by one or a few outliers[2]and

[2]Outliers and anomalies are essentially the same thing, and we make no formal distinction between them. But informally we think of anomalies as rare nuggets deserving of further analysis, while outliers are nuisance samples that contaminate the data of interest.

robust estimates of covariance [37]–[39] exclude or suppress the outliers in the computation. Excluding outliers is advisable in global estimates of covariance as well [40], [41].

2) Regression framework: By estimating what the target-free radiance *should* be at a pixel, we have a point of comparison with what the measured value of that pixel actually happens to be. It is common to make this estimate using the mean of pixels in an annulus around the pixel of interest. But there is more information in the annulus than this mean value [42], [43], and that suggests using more general estimators than just the mean. The derivation in [44] uses multivariate regression of the central pixel against the pixels in the surrounding annulus. This can be done on a band-by-band basis, or with multiple bands simultaneously. Other variants use median instead of mean [45], and a patch-based nearest neighbors regressor was also developed [46].

It is worth noting that the problem of estimating a central pixel, using the specific pixels that immediately surround it, along with the statistical context of the rest of the image, is a special case of the "inpainting" problem. In this case, only a single pixel is being inpainted at a time, but this inpainting is done for every pixel in the image.

II. Evaluation

A conceptual problem with anomaly detection is the ambiguous nature of "interestingness." But a more technical problem is that, since anomalies are by definition rare, it is difficult to find enough of them to do statistical comparisons of algorithms, and it is easy to be misdirected by anecdotal results.

Section I-B described how to resolve the ambiguity issue by treating anomalies as samples from a specific and well-defined, yet very broad and flat (*e.g.*, uniform), distribution. This resolution led, for instance in Eq. (1), to likelihood-based algorithms for anomaly detection.

An arguably more important advantage to treating anomalies as samples from a distribution is that it leads to a direct quantitative way to measure the performance of an anomaly detection algorithm, without relying on anecdotally identified anomalies in a given scene. The most straightforward way of exploiting this model is to use it to create artificial targets that can be "implanted" into the scene [26], [47]. This is usually performed at a judiciously chosen subset of locations, to avoid contaminating the background estimation with an unrealistically large number of targets. But in some cases one can more efficiently place a target at effectively *every* location in the image, producing matched pairs of with-target and without-target pixels from which to learn a target detector [48].

An advantage of explicit implanting is that it can be made as realistic as the simulation will allow. A potential disadvantage is that this explicit simulation can be expensive, may require *ad hoc* choices. Particularly for anomaly detection, where physical properties of the target are generally not well known, or at least not well specified, more generic approaches may be desirable.

With anomalies drawn from a uniform distribution, the volume inside a contour of anomalousness corresponds to undetected anomalies. Thus, a small volume is a proxy for a low missed detection rate; for a given false alarm rate, smaller volumes imply better anomaly detectors (Fig. 2). In an early example of this principle, Tax and Duin [49] used a uniform distribution of points to estimate volumes produced by different choice of kernel parameters, thereby providing a way to choose parameters without ground truth. Steinwart *et al.* [15] formalize the classification framework for anomaly detection and recommend a two-class classifier (a support vector machine, specifically) that distinguishes measured data from artificially generated uniformly distributed random samples. (See also Hastie *et al.* [50].)

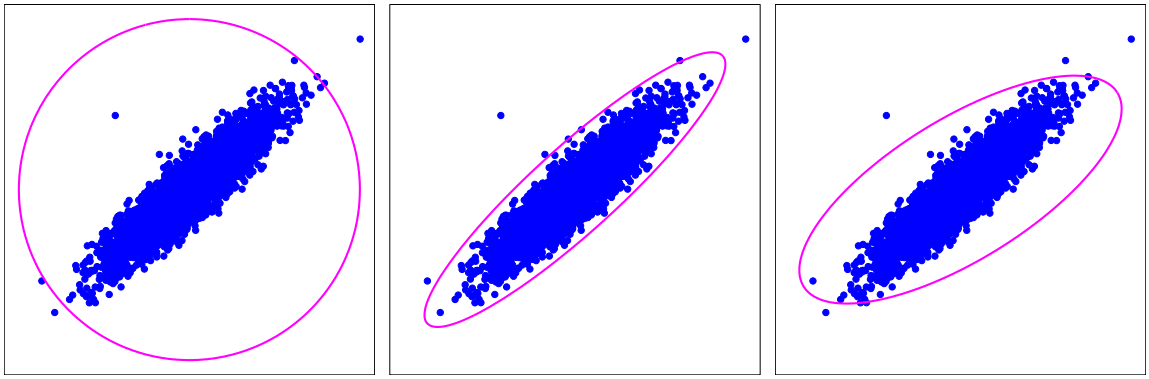


Fig. 2. Three anomaly detectors with the same false alarm rate ($P_{\text{fa}} = 0.001$). The contour marks the boundary between what is normal (inside) and what is anomalous (outside). If anomalies are presumed to be uniformly distributed, then the anomaly detector with smallest volume will have the fewest missed detections.

Plotting volume (or log-volume, which is often more convenient, especially in high dimensions) against false alarm rate provides a ROC-like curve that characterizes the anomaly detector's performance [51], [52].

For a covariance matrix, the volume is proportional to the determinant of the matrix. For local anomaly detectors described in Section I-C1, one can still use a global covariance based on the difference between measured and estimated (*e.g.*, by local mean) values at each pixel, and the smaller that covariance, the better the estimator. A natural choice, from a signal processing perspective is the total variance of that difference,

$$\sum_n (\mathbf{x}_n - \widehat{\mathbf{x}}_n)^T (\mathbf{x}_n - \widehat{\mathbf{x}}_n) \tag{6}$$

which corresponds to the trace of the covariance matrix. Smaller values of this variance imply that $\widehat{\mathbf{x}}$ is closer to $\mathbf{x}$, but Hasson *et al.* [45] point out that, in terms of target and anomaly detection performance, closer is not necessarily better.

When, instead of a global covariance estimator, we use a separate covariance for each pixel, based on the local neighborhood of that pixel, then it is more complicated. It is clear that the volumes of the individual covariances should be small, but it is not obvious how best to combine them. In Bachega *et al.* [53], it is argued, more on practical than theoretical grounds, that an average of the log volume is a good choice.

III. Periphery

For anomaly detection, low false alarm rates are imperative. So the challenge is to characterize the background density in regions where the data are sparse; that is, on the periphery (or "tail") of the distribution. Unfortunately, traditional density estimation methods, especially parametric estimators (*e.g.*, Gaussian), are dominated by the high-density core. And it bears pointing out that "robust" estimation methods (*e.g.*, [37], [54], [55]) achieve their robustness by paying even less attention to the periphery.

Robustness to outliers can be achieved by essentially removing the outliers from the data set. This direct approach is taken by the MCD (Minimum Covariance Determinant) of Rousseeuw *et al.* [37], [55]. For a data set of N samples, the idea is to take a core subset $\mathcal{H}$ of $h < N$ samples and to compute the mean and covariance from just the samples in $\mathcal{H}$, ignoring the rest. The

More attention to periphery →	More robust against outliers ↓	
	All data samples	Subset of h samples
Sample covariance matrix	Mahalanobis/RX	MCD
Minimum volume matrix	MVEE	MVEE-h

Fig. 3. Four algorithms for estimating ellipsoidal contours. All four algorithms seek ellipsoidal contours for the data, and can all four be expressed with an equation of the form $\mathcal{A}(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^T R^{-1}(\mathbf{x}-\boldsymbol{\mu})$. The top two algorithms use sample mean and sample covariance to estimate $\boldsymbol{\mu}$ and R, respectively; the bottom two seek a minimum volume ellipsoid that strictly encloses the data. The left two algorithms use all of the data in the training set; the right two algorithms use a subset $\mathcal{H}$ that includes *almost* all of the data. Note that the MVEE-h algorithm [41] is both robust to outliers and sensitive to data on the periphery of the distribution.

formal aim is to choose the subset so as to minimize the volume of the ellipsoid corresponding to the sample covariance. Specifically,

$$\begin{aligned} \min_{\mathcal{H}} \det(R) \quad \text{where} \quad & R = (1/h) \sum_{\mathbf{x}_n \in \mathcal{H}} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \\ \text{and} \quad & \boldsymbol{\mu} = (1/h) \sum_{\mathbf{x}_n \in \mathcal{H}} \mathbf{x}_n \\ \text{and} \quad & \#\{\mathcal{H}\} \geq h \end{aligned} \tag{7}$$

As stated, this is an NP-hard problem, but an iterative approach can be employed to find an approximate optimum. Given an initial set of core samples $\mathcal{H}$, we can compute $\boldsymbol{\mu}$ and R as sample mean and covariance of the core set. With this $\boldsymbol{\mu}$ and R, we can use Eq. (4) to compute $\mathcal{A}(\mathbf{x})$ for *all* of the samples. Taking the h samples with smallest $\mathcal{A}(\mathbf{x})$ values yields a new core set $\mathcal{H}'$. This process can be iterated, and is guaranteed to converge, though it is not guaranteed to converge to the global optimum defined in Eq. (7). Various tricks can be used both to speed up the iterations and to achieve lower minima [55].

Where MCD concentrates on identifying the core, the Minimum Volume Enclosing Ellipsoid (MVEE) algorithm concentrates on the periphery of the data. In contrast to Eq. (7), the aim is to optimize

$$\min_{\boldsymbol{\mu}, R} \det(R) \quad \text{where} \quad (\mathbf{x}_n - \boldsymbol{\mu})^T R^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) \leq 1 \quad \text{for all } n \tag{8}$$

Unlike the optimization in Eq. (7), the optimization here is convex and can be efficiently performed, using Khachiyan's algorithm [56], possibly including some of the further improvements that have since been suggested [57], [58].

Although robustness against outliers and sensitivity to the periphery are seemingly opposite requirements, practical anomaly detection actually wants both. For data sets in which a very small number of samples are truly outliers (or are truly anomalies), we *do* not want to include these samples in our characterization of the background. But absent these outliers, we do want to identify where the tail of background distribution is, and that requires attention to the samples on the periphery [51].

Fig. 3 illustrates this tension between attention to the periphery and robustness to outliers by showing four algorithms lined up along two axes. All four algorithms seek ellipsoidal contours

for the data, and all four can be expressed with an equation of the form in Eq. (4); what differs is the methodology for estimating $\boldsymbol{\mu}$ and R. As Fig. 3 suggests, it is possible to combine MCD and MVEE to create a new algorithm, called MVEE-h, that identifies a minimum volume ellipsoid that fully encloses not all, but most (in particular, a subset $\mathcal{H}$) of the data [41].

Other approaches have also been suggested for modifying the sample covariance in a way that better respects the points on the periphery. One family of such approaches uses the sample covariance to define eigenvectors, but modifies the eigenvalues in each of these eigenvector directions. This is consistent with the observation, noted by several authors [59]–[61], that tails tend to be heavier in some directions than others. The approach taken by Adler-Golden [61] was based on the observation that the heavy-tailed distribution has different properties in different eigenvector directions. A model is fit for each of these directions (but it's only one-dimensional so the fit is generally stable and robust), and this leads to a modified anomaly detector:

$$\mathcal{A}(\mathbf{x}) = \sum_i \left(a_i |x_{(i)}|\right)^{p_i}, \tag{9}$$

where $x_{(i)} = \lambda_i^{-1/2} \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$ is the ith whitened coordinate of $\mathbf{x}$. Here, where $\mathbf{u}_i$ and λ_i are the ith eigenvector and eigenvalue respectively, and a_i and p_i are obtained from the model fit for each component. If $a_i = 1$ and $p_i = 2$, then this is equivalent to Eq. (4). A related approach has also been suggested, in which $\mathcal{A}(\mathbf{x}) = \sum_i x_{(i)}^2 / \sigma_i^2$ and σ_i is based on inter-percentile difference instead of variance [51].

IV. Subspace

For many data sets, the covariance matrix usually exhibits a wide range of eigenvalues, and the smallest values correspond to directions that can be projected out of the data with minimal loss in accuracy. Indeed, it is often the case that data can be accurately represented by a lower-dimensional plane (or manifold [62]–[66]), and there are often advantages to analyzing the data in that lower-dimensional space. For anomaly detection, however, one has to be extra careful.

Qualitatively speaking, there are two kinds of anomalies: "in-plane" anomalies and "out-of-plane" anomalies. The in-plane anomalies are unusual with respect to the distribution that is projected onto the lower-dimensional high-variance subspace ("the plane"), and thus these tend to be large magnitude samples. The out-of-plane anomalies are unusual in that they are far from the plane, where "far" refers to distances larger than the smallest eigenvalues. Thus an out-of-plane anomaly can be a lower magnitude sample, and if the data sample is projected into the subspace, it may lose its anomalousness.

A simple measure of out-of-plane anomalousness is Euclidean distance to the subspace [67] (or, in more sophisticated cases, to the manifold [68]). The subspace RX (SSRX) algorithm effectively computes a Mahalanobis distance to the subspace [21]. In practice this is achieved by projecting to a dual subspace (that is, projecting *out* the high-variance directions) and then performing standard RX in that space.

The qualitative difference between the high-variance and low-variance directions has led to a variety of Gaussian/Non-Gaussian (G/NG) models for high-dimensional distributions. In these models, the low variance directions are modeled as Gaussian, but the high variance directions are modeled with simplex-based [52] or histogram-based [69] distributions. This enables more sophisticated models to be employed, but because they are only used for a few high-variance directions, the complexity is bounded, and the curse of dimensionality is ameliorated.

Another approach based on projection to a lower dimensional space was proposed by Kwon *et al.* [70]; here the projection operator is based on eigenvalues of a matrix that is the difference of two covariance matrices, one computed from an inner window (centered at the pixel under test in a moving window scenario) and one from an outer window (an annulus that surrounds the inner window and provides local context).

V. Kernels

A. Kernel density estimation

Given that the aim of anomaly detection is to estimate the background distribution $p_b(\mathbf{x})$, one of the most straightforward estimators is the kernel density estimator, or Parzen windows [71] estimator:

$$p_b(x) = (1/N) \sum_{n=1}^{N} \kappa(\mathbf{x}, \mathbf{x}_n), \tag{10}$$

where the sum is over all points in the data set, and where κ is a kernel function that is integrable and is everywhere non-negative. A popular choice is the Gaussian radial basis kernel,

$$\kappa(\mathbf{x}, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^d}} \exp\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right). \tag{11}$$

Eq. (11) requires the user to choose a "bandwidth" σ that characterizes, in some sense, the range of influence of each point. Since density can vary widely over a distribution, variable and data-adaptive bandwidth schemes have been proposed [72], [73].

In the limit as bandwidth goes to zero, the anomalousness at $\mathbf{x}$ is dominated by the $\kappa(\mathbf{x}, \mathbf{x}_i)$ associated with the $\mathbf{x}_i$ that is closest to $\mathbf{x}$. Indeed, the anomalousness in that case is equivalent to that distance. An anomaly detector based on distance to the nearest point has been proposed [74], though with an additional step that uses a graph-based approach to eliminate a small fraction (typically 5%) of the points to be used as $\mathbf{x}_i$. An updated variant was later proposed [75] that included normalization, subsampling, and a distance defined by the average of the distances to the third, fourth, and fifth nearest points.

B. Feature space interpretation: the "kernel trick"

A particularly fruitful (if initially counter-intuitive) interpretation of kernel functions is as dot products in a (usually higher-dimensional) feature space.

Let $\phi(\mathbf{x})$ be a function that maps $\mathbf{x}$ to some some feature space. Typically ϕ is nonlinear, and the map is to a feature space that is of higher dimension than $\mathbf{x}$. Scalar dot product in this feature space can be expressed as (again, typically nonlinear) functions of the values in the original data space. That is:

$$\kappa(\mathbf{r}, \mathbf{s}) = \phi(\mathbf{r})^T \phi(\mathbf{s}). \tag{12}$$

The "kernel trick" is the observation that even though the function ϕ and the feature space are presumed to "exist" in some abstract mathematical sense, we do not actually need to use ϕ, as long as we have the kernel function κ. A popular choice is the Gaussian kernel

$$\kappa(\mathbf{r}, \mathbf{s}) = \exp\left(-\frac{\|\mathbf{r} - \mathbf{s}\|^2}{2\sigma^2}\right), \tag{13}$$

but many options are available. Polynomial kernels, for example, are of the form $\kappa(\mathbf{r}, \mathbf{s}) = (c + \mathbf{r}^T\mathbf{s})^d$ for some polynomial dimension d. More general radial-basis kernels are scalar functions of the scalar value $\|\mathbf{r} - \mathbf{s}\|^2$; functions that are more heavy-tailed than the Gaussian have been proposed for this purpose [76].

This enables us re-derive the Parzen window detector from a different point of view. Given our data, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we first map to the feature space: $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)\}$. In this feature space we define the centroid

$$\boldsymbol{\mu}_\phi = \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n) \tag{14}$$

and we define anomalousness as distance to the centroid in this feature space.

$$\begin{aligned}\mathcal{A}(\mathbf{x}) &= \|\phi(\mathbf{x}) - \boldsymbol{\mu}_\phi\|^2 \\ &= (\phi(\mathbf{x}) - \boldsymbol{\mu}_\phi)^T(\phi(\mathbf{x}) - \boldsymbol{\mu}_\phi) \\ &= \phi(\mathbf{x})^T\phi(\mathbf{x}) - 2\phi(\mathbf{x})^T\boldsymbol{\mu}_\phi + \boldsymbol{\mu}_\phi^T\boldsymbol{\mu}_\phi \end{aligned} \tag{15}$$

We observe that first term $\phi(\mathbf{x})^T\phi(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x})$ is constant for radial basis kernels, that the third term is also constant, and that

$$\phi(\mathbf{x})^T\boldsymbol{\mu}_\phi = \frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x})^T\phi(\mathbf{x}_n) = \frac{1}{N}\sum_{n=1}^{N}\kappa(\mathbf{x}, \mathbf{x}_n). \tag{16}$$

This leads to

$$\mathcal{A}(\mathbf{x}) = \text{constant} - \frac{2}{N}\sum_{n=1}^{N}\kappa(\mathbf{x}, \mathbf{x}_n), \tag{17}$$

which is a negative monotonic transform of the density estimator $(1/N)\sum_{n=1}^{N}\kappa(\mathbf{x}, \mathbf{x}_n)$, and therefore equivalent to anomaly detection based on Parzen windows density estimation. The power of kernels in this case is that a seemingly trivial anomaly detector (Euclidean distance to the centroid of the data) in feature space maps back to a more complex data-adaptive anomaly detector in the data space.

The power of this feature-space interpretation of kernels is that it enables us to derive other expressions for anomaly detection, starting with very simple models in feature space that are then mapped back to more sophisticated data-adaptive anomaly detectors in the data space.

For instance, instead of Euclidean distance to the centroid $\boldsymbol{\mu}_\phi$, consider a more periphery-respecting model that uses an adaptive center $\mathbf{a}_\phi$ that is adjusted to minimize the radius of the sphere that encloses all of the data (see Fig. 4). That is,[3]

$$\min_{r, \mathbf{a}_\phi} r^2 \quad \text{subject to: } \|\phi(\mathbf{x}_n) - \mathbf{a}_\phi\|^2 \leq r^2 \tag{18}$$

or more generally, that *mostly* encloses the data:

$$\min_{r, \mathbf{a}_\phi, \xi} r^2 + c\sum_n \xi_n \tag{19}$$

$$\text{subject to: } \|\phi(\mathbf{x}_n) - \mathbf{a}_\phi\|^2 \leq r^2 + \xi_n \tag{20}$$

$$\text{and: } \xi_n \geq 0, \tag{21}$$

[3]Another way of expressing the centroid $\boldsymbol{\mu}_\phi$ is as the solution to the minimization of the average squared radius: $\boldsymbol{\mu}_\phi = \text{argmin}_{\boldsymbol{\mu}}\sum_n \|\phi(\mathbf{x}_n) - \boldsymbol{\mu}\|^2$; by comparison, we can say $\mathbf{a}_\phi$ is the solution to the minimization of the maximum squared radius: $\mathbf{a}_\phi = \text{argmin}_{\mathbf{a}}\text{max}_n\|\phi(\mathbf{x}_n) - \mathbf{a}\|^2$. We can interpret Eq. (19) as the minimization of a "soft" maximum.

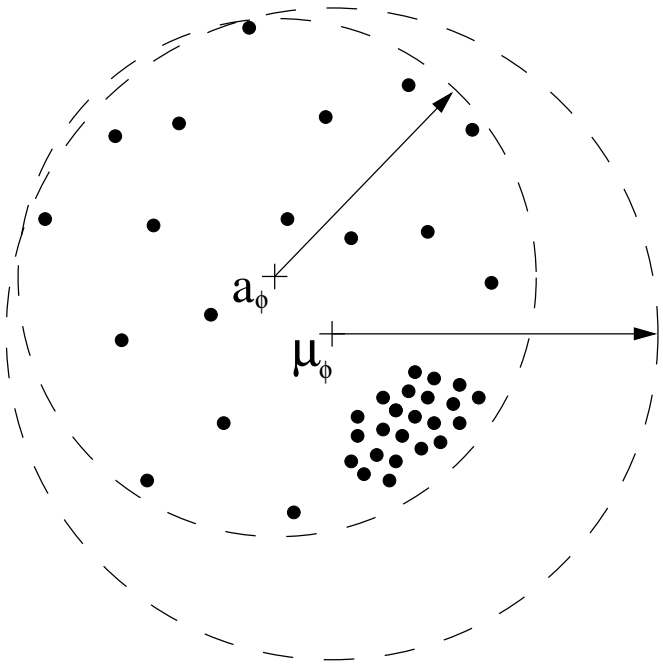


Fig. 4. Adaptive center versus data centroid. Data samples in feature space are indicated by black dots. Here, $\boldsymbol{\mu}_\phi = (1/N)\sum_{n=1}^{N}\phi(\mathbf{x}_n)$ is the centroid of the data, and a_ϕ is the adaptive center that enables the data to be enclosed by a smaller circle.

which is equivalent to Eq. (18) in the large c limit (which forces the "slack" variables ξ_n to zero). This optimization leads to the Support Vector Domain Decomposition (SVDD) anomaly detector [77], [78], which has the form

$$\mathcal{A}(\mathbf{x}) = \|\phi(\mathbf{x}) - \mathbf{a}_\phi\|^2 = \text{constant} - \sum_n a_n \kappa(\mathbf{x}, \mathbf{x}_n) \tag{22}$$

where the scalar coefficients a_n are positive and sum to 1. This is very much like the kernel density estimator for anomaly detection in Eq. (17), but it puts uneven weight on the points in the dataset. The SVDD is very similar to the support vector machine for one-class classification [9], and has the property that $a_n = 0$ for points deep in the interior of the distribution.

In keeping with the general strategy of mapping data to kernel space, and applying anomaly detection in this kernel space, we can also "kernelize" the RX algorithm. Here, given the data is mapped to kernel space $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)\}$, we need to compute a mean and covariance matrix. The mean $\boldsymbol{\mu}_\phi$ is defined in Eq. (14); to compute the covariance, we write

$$R_\phi = \sum_n [\phi(\mathbf{x}_n) - \boldsymbol{\mu}_\phi][\phi(\mathbf{x}_n) - \boldsymbol{\mu}_\phi]^T \tag{23}$$

A key step in the computation of RX anomalousness is the inversion of this covariance matrix. But R_ϕ has bounded rank (it is at most $n-1$), and depending on the dimension of the feature space, may not be invertible.

In Cremers *et al.* [79], this problem was addressed by regularizing to covariance, so that that $R_\phi + \lambda I$ was inverted, where λ was taken to be a small but nonzero value. In Kwon and

Nasrabadi [80], the pseudoinverse was taken. The effect of the pseudoinverse is to project data (in the feature space) to the in-sample data plane, but this projection can be problematic for anomaly detection [81]. Anomalies are different from the rest of the data, and this difference will be suppressed by projection back into the in-sample data plane.

Indeed, another kernelization that can be effective is the kernel subspace anomaly detector [82], [83]. Here, principal components analysis is performed in the feature space, and a subspace is defined that includes the first few principal components. Anomalousness is defined in terms of the distance to this subspace.

VI. Change

For the anomalous *change* detection problem, the aim is to find interesting differences between two images, taken of the same scene, but at different times and typically under different viewing conditions [84]. There will be some differences that are pervasive – *e.g.*, differences due to overall calibration, contrast, illumination, look-angle, focus, spatial misregistration, atmospheric or even seasonal changes – but there may also be changes that occur in only a few pixels. These rare changes potentially indicate something truly changed in the scene, and the idea is to use anomaly detection to find them. But our interest is in pixels where *changes* between the pixels are unusual, not so much in unusual pixels that are "similarly unusual" in both images. Informally speaking, we want to learn the "patterns" of these pervasive differences, and then the changes that do not fit the patterns are identified as anomalous.

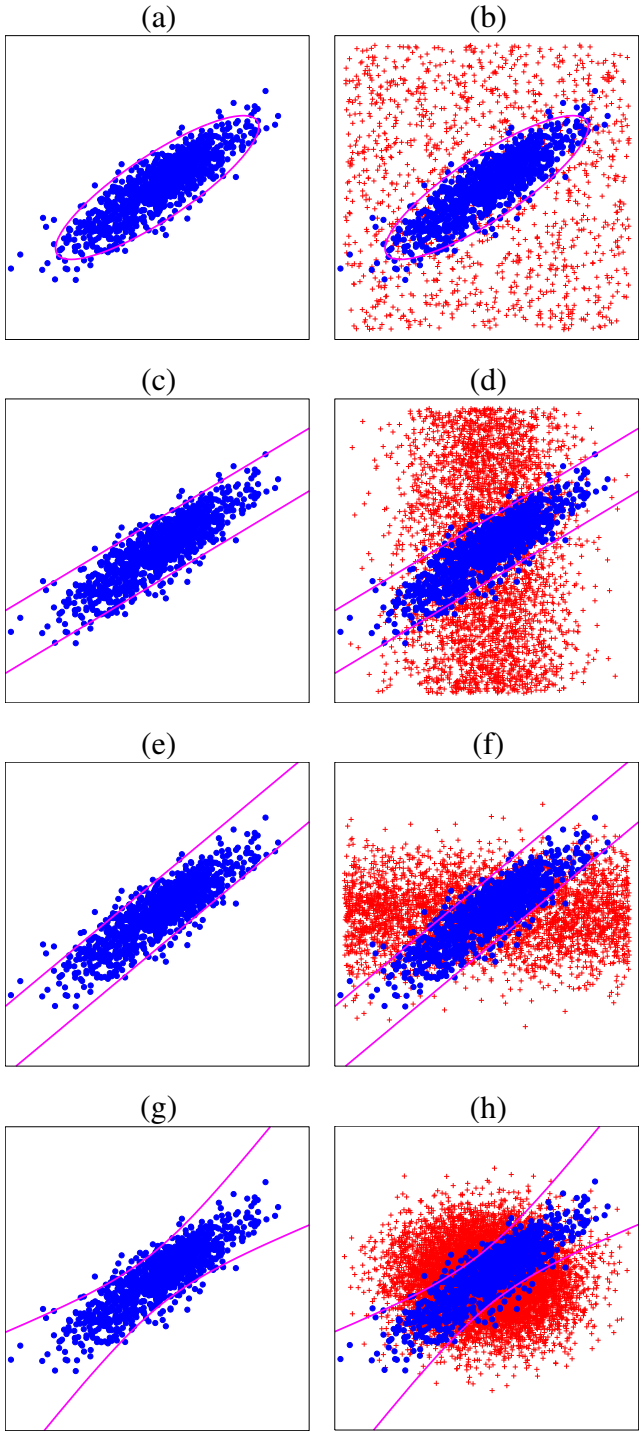
An important precursor to anomalous change detection is the co-registration of the two images. We say that images are registered if corresponding pixels in the two images correspond to the same position in the scene. Registering imagery is a nontrivial task, yet misregistration is one of the main confounds in change detection [85]–[88]. In what follows, let $\mathbf{x}$ and $\mathbf{y}$ refer to corresponding pixels in two images.

A. Subtraction-based approaches to anomalous change detection

The most straightforward way to look for changes in a pair of images is to subtract them, $\mathbf{e} = \mathbf{y} - \mathbf{x}$, and then to restrict analysis to the difference image $\mathbf{e}$ [89]. Simple subtraction, although it has the advantage of being simple, has the disadvantage that it folds in pervasive differences along with the anomalous changes.

Most anomalous change detection algorithms are based on subtracting images, but involve transforming the images to make them more similar. For instance, the chronochrome [90] seeks a linear transform of the first image to make it as similar as possible (in a least squares sense) to the second image. That is, it seeks L so that $\|\mathbf{y} - L\mathbf{x}\|^2$, averaged over the whole image, is minimized. To simplify notation, we will assume means have been subtracted from $\mathbf{x}$ and $\mathbf{y}$, and define the covariance matrices $X = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle$, $Y = \left\langle \mathbf{y}\mathbf{y}^T \right\rangle$, and $C = \left\langle \mathbf{y}\mathbf{x}^T \right\rangle$. The linear transform that minimizes the least square fit of $\mathbf{y}$ to $L\mathbf{x}$ is given by $L = CX^{-1}$. Now the subtraction that is performed is $\mathbf{e} = \mathbf{y} - L\mathbf{x}$, and this reduces the effect of pervasive differences on $\mathbf{e}$ while still "letting through" the anomalous changes. Note that there is an asymmetry in the chronochrome; by swapping the role of $\mathbf{x}$ and $\mathbf{y}$, and seeking L' to minimize $\|\mathbf{e} = \mathbf{x} - L'\mathbf{y}\|^2$, one obtains a different anomalous change detector. Clifton [91] proposed a neural network version of chronochrome, in a nonlinear function $\mathcal{L}(\mathbf{x})$ is chosen to minimize $\mathbf{e} = \mathbf{y} - \mathcal{L}(\mathbf{x})$, with the aim of even further suppressing the pervasive differences.

A more symmetrical approach, which is sometimes called covariance equalization [92], [93] or whitening/de-whitening [94], transforms the data in both images before it subtracts them:



$$p_a(\mathbf{x}, \mathbf{y}) = u(\mathbf{x})u(\mathbf{y})$$

$$p_a(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})u(\mathbf{y})$$

$$p_a(\mathbf{x}, \mathbf{y}) = u(\mathbf{x})p(\mathbf{y})$$

$$p_a(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

Fig. 5. Four anomalous change detectors, derived from four distinct and explicit definitions of anomalous change. Here $\mathbf{x}$ is represented by the horizontal axis and $\mathbf{y}$ is the vertical axis. The left panels show the non-anomalous data (sampled from a correlated Gaussian distribution) and the boundaries outside of which are the points that the detectors consider anomalous changes. The panels to the right include samples drawn from $p_a(\mathbf{x}, \mathbf{y})$, the model for the distribution of anomalous changes. (a,b) RX detector is obtained from "straight" anomaly detection; (c,d) Chronochrome detector optimized for $\mathbf{x} \rightarrow \mathbf{y}$ changes; (e,f) Chronochrome optimized for $\mathbf{y} \rightarrow \mathbf{x}$ changes; and (g,h) hyperbolic anomalous change detector.

$\mathbf{e} = Y^{-1/2}\mathbf{y} - X^{-1/2}\mathbf{x}$. It can be shown [95] that this is related to canonical coordinate analysis and to Nielsen's multivariate alteration detection (MAD) algorithm [96] and to a "total least squares" change detection algorithm [97]. One reason there are so many variants is that there is not a unique whitening transform: if U is an arbitrary orthogonal matrix, then $UX^{-1/2}$ will also whiten the $\mathbf{x}$ samples. That is: if $\widehat{\mathbf{x}} = UX^{-1/2}\mathbf{x}$, then we say that $\widehat{\mathbf{x}}$ is whitened because $\left\langle \widehat{\mathbf{x}}\widehat{\mathbf{x}}^T \right\rangle = I$.

In all of these algorithms, a vector-valued difference, $\mathbf{e}$, is produced. For anomaly detection, we need a scalar valued measure of anomalousness, and the RX formula provides the most straightforward way to achieve that. Let $E = \left\langle \mathbf{e}\mathbf{e}^T \right\rangle$ be the covariance matrix of the residuals, and then take the anomalousness to be

$$\mathcal{A}(\mathbf{e}) = \mathbf{e}^T E^{-1} \mathbf{e} \tag{24}$$

As a particular example, we can write Eq. (24) for the chronochrome detector. Here $\mathbf{e} = y - CX^{-1}\mathbf{x}$, and

$$\mathcal{A}(\mathbf{x}, \mathbf{y}) = (y - CX^{-1}\mathbf{x})^T [Y - CX^{-1}C]^{-1} (y - CX^{-1}\mathbf{x}) \tag{25}$$

B. Distribution-based approaches to anomalous change detection

While subtracting (suitably transformed) images is an intuitively plausible way to look for anomalous changes, we can take more of a machine learning point of view and treat the problem as one of two-class classification, where the two classes are pervasive differences and anomalous changes. If we can write down expressions for the underlying distributions of these two classes, then their ratio will be an optimal detector of anomalous changes.

The distribution for the pervasive differences is just the distribution of the data itself. We can call this $p_b(\mathbf{x}, \mathbf{y})$ to indicate that it is the "background" distribution. What we need is a generative model for anomalous changes; that is, a distribution $p_a(\mathbf{x}, \mathbf{y})$ that describes what we mean when we speak of anomalous changes. What that in hand, our anomaly detector is given by the likelihood ratio

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \frac{p_a(\mathbf{x}, \mathbf{y})}{p_b(\mathbf{x}, \mathbf{y})}. \tag{26}$$

The simplest choice, following our experience with straight anomaly detection, is a uniform distribution. We can write this $p_a(\mathbf{x}, \mathbf{y}) = u(\mathbf{x})u(\mathbf{y})$. This is not an unreasonable choice, but it does not put any particular emphasis on *change*. A pair of pixels $(\mathbf{x}, \mathbf{y})$ that are "similarly anomalous" in both images (*e.g.*, the pixel might be unusually bright in both images) do not particularly indicate that something has changed at that pixel position in the scene.

But there are alternative models $p_a(\mathbf{x}, \mathbf{y})$ that correspond to different notions of what an anomalous change is. For instance, if we are thinking specifically of changes $\mathbf{x} \rightarrow \mathbf{y}$, where $\mathbf{x}$ is a kind of "reference" image, and $\mathbf{y}$ is the new image that might contain the changes of interest, then $p_a(\mathbf{x}, \mathbf{y}) = p_b(\mathbf{x})u(\mathbf{y})$. This corresponds to a non-anomalous $\mathbf{x}$ and an anomalous $\mathbf{y}$. It is interesting to note that using this definition of anomaly in Eq. (26) leads to $\mathcal{L}(\mathbf{x}, \mathbf{y}) = 1/p_b(\mathbf{y}|\mathbf{x})$, so that anomalousness of $\mathbf{y}$ varies inversely with the *conditional* probability density of $\mathbf{y}$. As with the chronochrome[4], there is an asymmetry in this model of anomalous change; its mirror image is the situation in which $\mathbf{y}$ is considered the reference image and it is $\mathbf{y} \rightarrow \mathbf{x}$ changes that are of particular interest; here, $p_a(\mathbf{x}, \mathbf{y}) = u(\mathbf{x})p_b(\mathbf{y})$.

[4]In fact, the chronochrome can be obtained from this model in the special case that $p_b(\mathbf{x}, \mathbf{y})$ is Gaussian [98].

Another choice for $p_a(\mathbf{x}, \mathbf{y})$ has also been suggested [99]. Here, $p_a(\mathbf{x}, \mathbf{y}) = p_b(\mathbf{x})p_b(\mathbf{y})$, and $\mathbf{x} \to \mathbf{y}$ and $\mathbf{y} \to \mathbf{x}$ changes are treated equally. The informal interpretation is that unusual changes are pixel pairs $(\mathbf{x}, \mathbf{y})$ which are collectively unusual, but individually normal. That is, the $\mathbf{x}$ pixel value is typical for the x-image, and the $\mathbf{y}$ pixel value is typical for the y-image, but the $(\mathbf{x}, \mathbf{y})$ pair is unusual. If we use this model for anomalies in Eq. (26), and take a logarithm, we obtain

$$\log \mathcal{L}(\mathbf{x}, \mathbf{y}) = \log p_b(\mathbf{x}) + \log p_b(\mathbf{y}) - \log p_b(\mathbf{x}, \mathbf{y}) \tag{27}$$

an expression for anomalousness that looks like negative mutual information of $\mathbf{x}$ and $\mathbf{y}$.

In the case of Gaussian $p_b(\mathbf{x}, \mathbf{y})$, Eq. (27) reduces to a quadratic expression in $\mathbf{x}$ and $\mathbf{y}$ that has hyperbolic contours (see Fig. 5(g,h)). Experiments with real and simulated anomalous changes in real imagery indicated that this hyperbolic anomalous change detection (HACD) generally outperformed the subtraction-based anomaly detectors [95].

A further advantage of the distribution-based approach is that the distribution needn't be Gaussian. Indeed, we can take a purely nonparametric view, and treat the problem of distinguishing pervasive differences from anomalous changes as a machine learning classification. Steinwart *et al.* [100] used support vector machines for just this purpose. But a simpler approach, that has also proven effective, is to consider a parametric distribution, but one slightly more general than the Gaussian. The class of elliptically-contoured (EC) distributions are, like the Gaussian, primarily parameterized by a mean vector and covariance matrix, but do not share the sharp e^{-r^2} tail of the Gaussian. Heavy-tailed EC distributions have been suggested for hyperspectral imagery in general [101], and for anomalous change detection in particular [102], [103]. Although EC distributions do not affect the "straight anomaly detector" shown in Fig. 5(a,b), they do generalize chronochrome and hyperbolic anomalous change detectors in a way that can lead to improved performance [103]. Kernelization of the EC-based change detector has also been shown to be advantageous [104].

C. Further comments on anomalous change detection

The description here treats pixel pairs as independent samples from an unknown distribution. But there is a lot of spatial structure in imagery, and further gains can be made by incorporating spatial aspects along with the spectral [105], [106]. The description here also considers only pairs of images; often there are more than two images, and these algorithms can be extended to that case [107], [108], though this approach may not be optimal for sequences of images (*e.g.*, anomalous activities in video) where the order of the images in the sequence matters. Another issue that arises in remote sensing is that the anomalous targets may be subpixel in extent, which leads to a different optimization problem [109].

Anomalous change detection is a problem that is particularly well matched to remote sensing imagery, and in that context, a variety of practical issues have been discussed [110], [111]. One of the biggest of these issues is misregistration, when the images don't exactly line up (and they never *exactly* line up). Although the effects of misregistration can to some extent be learned from the pervasive differences it creates in image pairs, it is still one of the main confounds to change detection [85], [86]. Gains can be made by explicitly adapting the change detection algorithm to be more robust to misregistration error [87], [88].

VII. Conclusion

Anomaly detection is seldom a goal in its own right. It is the first step in a search for data samples that are relevant, meaningful, or – in some sense that depends on where the data came

from and what they are being used for – interesting.

The mystical "by definition undefined" aspect of anomaly detection mostly derives from the ambiguity of what one means by "interesting" and this has led to a wide variety of *ad hoc* anomaly detection algorithms, justified by hand-waving arguments and validated (if at all) by anecdotal performance on imagery with a statistically inadequate number of pre-judged anomalies.

By employing a framework in which anomalies are in fact well-defined, as samples drawn from some broad and flat distribution, anomaly detection algorithms can be objectively tested, and improvements can be confidently constructed.

Within this framework, many of the tools that have been developed for signal processing, machine learning, and data analytics in general, can be brought to bear on the detection of anomalies. These range from the venerable Gaussian distribution to kernels and subspaces (and kernelized subspaces!), and invoke the usual issues in underfitting and overfitting data.

The technical challenge of anomaly detection is not usually the anomalies themselves, but with characterizing what can be a complex and highly structured background. Since most anomaly detection scenarios require a low false alarm rate, it is out on the periphery of this background where the modeling is emphasized. This is something of a challenge, since the data density is much lower there. The modeling, however is discriminative, not generative, which means that the aim is not the model the distribution *per se*, but to find the boundary that separates the non-anomalous data from the anomalies.

REFERENCES

1. S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images." *IEEE A&E Systems Magazine* **25**, 5–27 (2010).
2. 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* **19**, 58–69 (Jan, 2002).
3. E. A. Ashton, "Multialgorithm solution for automated multispectral target detection." *Optical Engineering* **38**, 717–724 (1999).
4. B. A. Whitehead and W. A. Hoyt, "Function approximation approach to anomaly detection in propulsion system test data." *J. Propulsion and Power* **11**, 1074–1076 (1995).
5. K. Worden, "Structural fault detection using a novelty measure." *J. Sound and Vibration* **201**, 85–101 (1997).
6. C. Manikopoulos and S. Papavassiliou, "Network intrusion and fault detection: a statistical anomaly approach." *IEEE Communications Magazine* **40**, no. 10, 76–82 (2002).
7. M. Thottan and C. Ji, "Anomaly detection in IP networks." *IEEE Trans. Signal Processing* **51**, 2191–2204 (2003).
8. D. M. Cai, M. Gokhale, and J. Theiler, "Comparison of feature selection and classification algorithms in identifying malicious executables." *Computational Statistics and Data Analysis* **51**, 3156–3172 (2007).
9. B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection." *Advances in Neural Information Processing Systems* **12**, 582–588 (1999).
10. M. Markou and S. Singh, "Novelty detection: a review – part 1: statistical approaches." *Signal Processing* **83**, 2481–2497 (2003).
11. M. Markou and S. Singh, "Novelty detection: a review – part 2: neural network based approaches." *Signal Processing* **83**, 2499–2521 (2003).
12. D. M. Tax, *One-class classification: Concept-learning in the absence of counter-examples*. (TU Delft, Delft University of Technology, 2001).
13. L. M. Manevitz and M. Yousef, "One-class SVMs for document classification." *J. Machine Learning Res.* **2**, 139–154 (2001).
14. 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).

15. I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection." *J. Machine Learning Research* **6**, 211–232 (2005).
16. E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. (Springer, New York, 2005).
17. J. Theiler, "Confusion and clairvoyance: some remarks on the composite hypothesis testing problem." *Proc. SPIE* **8390**, 839003 (2012).
18. P. C. Mahalanobis, "On the generalised distance in statistics." *Proc. National Institute of Sciences of India* **2**, 49–55 (1936).
19. 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* **38**, 1760–1770 (1990).
20. A. D. Stocker, I. S. Reed, and X. Yu, "Multi-dimensional signal processing for electro-optical target detection." *Proc. SPIE* **1305**, 218–231 (1990).
21. A. Schaum, "Hyperspectral anomaly detection: Beyond RX." *Proc. SPIE* **6565**, 656502 (2007).
22. J. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images." *Proc. SPIE* **5093**, 230–240 (2003).
23. A. Hayden, E. Niple, and B. Boyce, "Determination of trace-gas amounts in plumes by the use of orthogonal digital filtering of thermal-emission spectra." *Applied Optics* **35**, 2802–2809 (1996).
24. J. Theiler and B. Wohlberg, "Detection of unknown gas-phase chemical plumes in hyperspectral imagery." *Proc. SPIE* **8743**, 874315 (2013).
25. 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* **7**, 2317–2336 (2014).
26. Y. Cohen, Y. August, D. G. Blumberg, and S. R. Rotman, "Evaluating sub-pixel target detection algorithms in hyper-spectral imagery." *J. Electrical and Computer Engineering* **2012**, 103286 (2012).
27. S. Matteoli, M. Diani, and G. Corsini, "Improved estimation of local background covariance matrix for anomaly detection in hyperspectral images." *Optical Engineering* **49**, 046201 (2010).
28. J. H. Friedman, "Regularized discriminant analysis." *J. Am. Statistical Assoc.* **84**, 165–175 (1989).
29. N. M. Nasrabadi, "Regularization for spectral matched filter and RX anomaly detector." *Proc. SPIE* **6966**, 696604 (2008).
30. J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data." *IEEE Trans. Pattern Analysis and Machine Intelligence* **18**, 763–767 (1996).
31. J. Theiler, "The incredible shrinking covariance estimator." *Proc. SPIE* **8391**, 83910P (2012).
32. G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform." *Advances in Neural Information Processing Systems* **21**, 225–232 (2009).
33. J. Theiler, G. Cao, L. R. Bachega, and C. A. Bouman, "Sparse matrix transform for hyperspectral image processing." *IEEE J. Selected Topics in Signal Processing* **5**, 424–437 (2011).
34. C. E. Caefer, J. Silverman, O. Orthal, D. Antonelli, Y. Sharoni, and S. R. Rotman, "Improved covariance matrices for point target detection in hyperspectral data." *Optical Engineering* **47**, 076402 (2008).
35. M. J. Carlotto, "A cluster-based approach for detecting man-made objects and changes in imagery." *IEEE Trans. Geoscience and Remote Sensing* **43**, 374–387 (2005).
36. J. Theiler and L. Prasad, "Overlapping image segmentation for context-dependent anomaly detection." *Proc. SPIE* **8048**, 804807 (2011).
37. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. (Wiley-Interscience, New York, 1987).
38. S. Matteoli, M. Diani, and G. Corsini, "Hyperspectral anomaly detection with kurtosis-driven local covariance matrix corruption mitigation." *IEEE Geoscience and Remote Sensing Lett.* **8**, 532–536 (2011).
39. S. Matteoli, M. Diani, and G. Corsini, "Impact of signal contamination on the adaptive detection performance of local hyperspectral anomalies." *IEEE Trans. Geoscience and Remote Sensing* **52**, 1948–1968 (2014).
40. W. F. Basener, "Clutter and anomaly removal for enhanced target detection." *Proc. SPIE* **7695**, 769525 (2010).
41. G. Grosklos and J. Theiler, "Ellipsoids for anomaly detection in remote sensing imagery." *Proc. SPIE* **9472**, 94720P (2015).
42. J. Theiler and J. Bloch, "Multiple concentric annuli for characterizing spatially nonuniform backgrounds." *The Astrophysical Journal* **519**, 372–388 (1999).
43. C. E. Caefer, M. S. Stefanou, E. D. Nelson, A. P. Rizzuto, O. Raviv, and S. R. Rotman, "Analysis of false alarm distributions in the development and evaluation of hyperspectral point target detection algorithms."

Optical Engineering **46**, 076402 (2007).
44. J. Theiler, "Symmetrized regression for hyperspectral background estimation." *Proc. SPIE* **9472**, 94721G (2015).
45. N. Hasson, S. Asulin, S. R. Rotman, and D. Blumberg, "Evaluating backgrounds for subpixel target detection: when closer isn't better." *Proc. SPIE* **9472**, 94720R (2015).
46. J. Theiler and B. Wohlberg, "Regression framework for background estimation in remote sensing imagery." *Proc. 5th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (2013).
47. W. F. Basener, E. Nance, and J. Kerekes, "The target implant method for predicting target difficulty and detector performance in hyperspectral imagery." *Proc. SPIE* **8048**, 80481H (2011).
48. J. Theiler, "Matched-pair machine learning." *Technometrics* **55**, 536–547 (2013).
49. D. Tax and R. Duin, "Uniform object generation for optimizing one-class classifiers." *J. Machine Learning Res.* **2**, 155–173 (2002).
50. T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer-Verlag, New York, 2001). This anomaly detection approach is developed in Chapter 14.2.4, and illustrated in Fig 14.3.
51. J. Theiler and D. Hush, "Statistics for characterizing data on the periphery." *Proc. IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)* 4764–4767 (2010).
52. J. Theiler, "Ellipsoid-simplex hybrid for hyperspectral anomaly detection." *Proc. 3rd IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (2011).
53. L. Bachega, J. Theiler, and C. A. Bouman, "Evaluating and improving local hyperspectral anomaly detectors." *IEEE Applied Imagery and Pattern Recognition (AIPR) Workshop* **39** (2011).
54. N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation." *Applied Statistics* **29**, 231–237 (1980).
55. P. J. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator." *Technometrics* **41**, 212–223 (1999).
56. L. G. Khachiyan, "Rounding of polytopes in the real number model of computation." *Mathematics of Operations Research* **21**, 307–320 (1996).
57. P. Kumar and E. A. Yildirim, "Minimum-volume enclosing ellipsoids and core sets." *J. Optimization Theory and Applications* **126**, 1–21 (2005).
58. M. J. Todd and E. A. Yildirim, "On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids." *Discrete Applied Mathematics* **155**, 1731–1744 (2007).
59. J. Theiler, B. R. Foy, and A. M. Fraser, "Characterizing non-Gaussian clutter and detecting weak gaseous plumes in hyperspectral imagery." *Proc. SPIE* **5806**, 182–193 (2005).
60. P. Bajorski, "Maximum Gaussianity models for hyperspectral images." *Proc. SPIE* **6966**, 69661M (2008).
61. S. M. Adler-Golden, "Improved hyperspectral anomaly detection in heavy-tailed backgrounds." *Proc. 1st IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (2009).
62. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding." *Science* **290**, 2323–2326 (2000).
63. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." *Science* **290**, 2319–2323 (2000).
64. C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery." *IEEE Trans. Geoscience and Remote Sensing* **43**, 441–454 (2005).
65. D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning." *IEEE Signal Processing Magazine* **31**, 55–66 (Jan, 2014).
66. A. K. Ziemann and D. W. Messinger, "An adaptive locally linear embedding manifold learning approach for hyperspectral target detection." *Proc. SPIE* **9472**, 94720O (2015).
67. K. Ranney and M. Soumekh, "Hyperspectral anomaly detection within the signal subspace." *IEEE Geoscience and Remote Sensing Letters* **3**, 312–316 (2006).
68. L. Ma, M. M. Crawford, and J. Tian, "Anomaly detection for hyperspectral images based on robust locally linear embedding." *J. Infrared Millimeter Terahertz Waves* **31**, 753–762 (2010).
69. G. A. Tidhar and S. R. Rotman, "Target detection in inhomogeneous non-Gaussian hyperspectral data based

on nonparametric density estimation." *Proc. SPIE* **8743**, 87431A (2013).
70. H. Kwon, S. Z. Der, and N. M. Nasrabadi, "Adaptive anomaly detection using subspace separation for hyperspectral imagery." *Optical Engineering* **42**, 3342–3351
71. E. Parzen, "On estimation of probability density function and mode." *Ann. Mathematical Statistics* **33**, 1065–1076 (1962).
72. 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 Geoscience and Remote Sensing Lett.* **11**, 163–167 (2014).
73. S. Matteoli, T. Veracini, M. Diani, and G. Corsini, "A locally adaptive background density estimator: An evolution for RX-based anomaly detectors." *IEEE Geoscience and Remote Sensing Lett.* **11**, 323–327 (2014).
74. B. Basener, E. Ientilucci, and D. Messinger, "Anomaly detection using topology." *Proc. SPIE* **6565**, 65650J (2007).
75. W. F. Basener and D. W. Messinger, "Enhanced detection and visualization of anomalies in spectral imagery." *Proc. SPIE* **7334**, 73341Q (2009).
76. C. Scovel, D. Hush, I. Steinwart, and J. Theiler, "Radial kernels and their reproducing kernel Hilbert spaces." *J. Complexity* **26**, 641–660 (2010).
77. D. Tax and R. Duin, "Data domain description by support vectors." In *Proc. ESANN99*, M. Verleysen, ed. (D. Facto Press, Brussels, 1999), pp. 251–256.
78. A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery." *IEEE Trans. Geoscience and Remote Sensing* **44**, 2282–2291 (2006).
79. D. Cremers, T. Kohlberger, and C. Schnörr, "Shape statistics in kernel space for variational image segmentation." *Pattern Recognition* **36**, 1929–1943 (2003).
80. H. Kwon and N. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery." *IEEE Trans. Geoscience and Remote Sensing* **43**, 388–397 (2005).
81. J. Theiler and G. Grosklos, "Problematic projection to the in-sample subspace for a kernelized anomaly detector." *IEEE Geoscience and Remote Sensing Lett.* (2016). To appear.
82. H. Hoffmann, "Kernel PCA for novelty detection." *Pattern Recognition* **40**, 863–874 (2007).
83. N. M. Nasrabadi, "Kernel subspace-based anomaly detection for hyperspectral imagery." *Proc. 1st IEEE Workshop on Hyperspectral Signal and Image Processing: Evolution in Remote Sensing (WHISPERS)* (2009).
84. A. Schaum and E. Allman, "Advanced algorithms for autonomous hyperspectral change detection." *IEEE Applied Imagery Pattern Recognition (AIPR) Workshop* **33**, 33–38 (2005).
85. J. Theiler, "Sensitivity of anomalous change detection to small misregistration errors." *Proc. SPIE* **6966**, 69660X (2008).
86. J. Meola and M. T. Eismann, "Image misregistration effects on hyperspectral change detection." *Proc. SPIE* **6966**, 69660Y (2008).
87. J. Theiler and B. Wohlberg, "Local co-registration adjustment for anomalous change detection." *IEEE Trans. Geoscience and Remote Sensing* **50**, 3107–3116 (2012).
88. K. Vongsy, M. T. Eismann, and M. J. Mendenhall, "Extension of the linear chromodynamics model for spectral change detection in the presence of residual spatial misregistration." *IEEE Trans. Geoscience and Remote Sensing* **53**, 3005–3021 (2015).
89. L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection." *IEEE Trans. Geoscience and Remote Sensing* **38**, 1171–1182 (2000).
90. A. Schaum and A. Stocker, "Long-interval chronochrome target detection." *Proc. ISSSR (Int. Symposium on Spectral Sensing Research)* (1998).
91. C. Clifton, "Change detection in overhead imagery using neural networks." *Applied Intelligence* **18**, 215–234 (2003).
92. A. Schaum and A. Stocker, "Linear chromodynamics models for hyperspectral target detection." *Proc. IEEE Aerospace Conference* 1879–1885 (2003).
93. A. Schaum and A. Stocker, "Hyperspectral change detection and supervised matched filtering based on covariance equalization." *Proc. SPIE* **5425**, 77–90 (2004).
94. R. Mayer, F. Bucholtz, and D. Scribner, "Object detection by using "whitening/dewhitening" to transform target signatures in multitemporal hyperspectral and multispectral imagery." *IEEE Trans. Geoscience and Remote Sensing* **41**, 1136–1142 (2003).
95. J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors." *Applied*

Optics **47**, F12–F26 (2008).
96. A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF post-processing in multispectral bi-temporal image data: new approaches to change detection studies." *Remote Sensing of the Environment* **64**, 1–19 (1998).
97. J. Theiler and A. Matsekh, "Total least squares for anomalous change detection." *Proc. SPIE* **7695**, 76951H (2010).
98. J. Theiler and S. Perkins, "Resampling approach for anomalous change detection." *Proc. SPIE* **6565**, 65651U (2007).
99. J. Theiler and S. Perkins, "Proposed framework for anomalous change detection." *ICML Workshop on Machine Learning Algorithms for Surveillance and Event Detection* 7–14 (2006).
100. I. Steinwart, J. Theiler, and D. Llamocca, "Using support vector machines for anomalous change detection." *Proc. IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)* 3732–3735 (2010).
101. D. Manolakis, D. Marden, J. Kerekes, and G. Shaw, "On the statistics of hyperspectral imaging data." *Proc. SPIE* **4381**, 308–316 (2001).
102. A. Schaum, E. Allman, J. Kershenstein, and D. Alexa, "Hyperspectral change detection in high clutter using elliptically contoured distributions." *Proc. SPIE* **6565**, 656515 (2007).
103. J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy, "Elliptically-contoured distributions for anomalous change detection in hyperspectral imagery." *IEEE Geoscience and Remote Sensing Lett.* **7**, 271–275 (2010).
104. N. Longbotham and G. Camps-Valls, "A family of kernel anomaly change detectors." *Proc. 6th IEEE Workshop on Hyperspectral Signal and Image Processing: Evolution in Remote Sensing (WHISPERS)* (2014).
105. T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on Markov random field models." *IEEE Trans. Geoscience and Remote Sensing* **40**, 1815–1823 (2002).
106. J. Theiler, "Spatio-spectral anomalous change detection in hyperspectral imagery." *Proc. 1st IEEE Global Signal and Information Processing Conference* 953–956 (2013).
107. S. M. Adler-Golden, S. C. Richtsmeier, and R. Shroll, "Suppression of subpixel sensor jitter fluctuations using temporal whitening." *Proc. SPIE* **6969**, 69691D (2008).
108. J. Theiler and S. M. Adler-Golden, "Detection of ephemeral changes in sequences of images." *IEEE Applied Imagery Pattern Recognition (AIPR) Workshop* **37** (2009).
109. J. Theiler, "Subpixel anomalous change detection in remote sensing imagery." *Proc. IEEE Southwest Symposium on Image Analysis and Interpretation* 165–168 (2008).
110. M. T. Eismann, J. Meola, and R. Hardie, "Hyperspectral change detection in the presence of diurnal and seasonal variations." *IEEE Trans. Geoscience and Remote Sensing* **46**, 237–249 (2008).
111. M. T. Eismann, J. Meola, A. D. Stocker, S. G. Beaven, and A. P. Schaum, "Airborne hyperspectral detection of small changes." *Applied Optics* **47**, F27–F45 (2008).