# Resampling approach for anomalous change detection

James Theiler[a] and Simon Perkins[b]

[a]Space and Remote Sensing Sciences Group
Los Alamos National Laboratory, Los Alamos, NM, USA.

[b]UltraSpectral, Inc. (http://www.ultraspectral.com)
5701 Carmel Ave NE, Albuquerque, NM, USA

## ABSTRACT

We investigate the problem of identifying pixels in pairs of co-registered images that correspond to real changes on the ground. Changes that are due to environmental differences (illumination, atmospheric distortion, etc.) or sensor differences (focus, contrast, etc.) will be widespread throughout the image, and the aim is to avoid these changes in favor of changes that occur in only one or a few pixels. Formal outlier detection schemes (such as the one-class support vector machine) can identify rare occurrences, but will be confounded by pixels that are "equally rare" in both images: they may be anomalous, but they are not changes. We describe a resampling scheme we have developed that formally addresses both of these issues, and reduces the problem to a binary classification, a problem for which a large variety of machine learning tools have been developed. In principle, the effects of misregistration will manifest themselves as pervasive changes, and our method will be robust against them – but in practice, misregistration remains a serious issue.

**Keywords:** remote sensing, change detection, anomaly detection, machine learning, support vector machine, image registration

## 1. INTRODUCTION: THE CASE FOR AUTOMATED CHANGE DETECTION

With decreasing costs of acquiring overhead imagery, and an increasing variety of available image modalities (from video to hyperspectral), image analysts have the ability to see more of the world in more detail than ever before. For precisely the same reasons, image analysts also have more images than they can look at, let alone analyze, which makes the automated analysis of imagery an imperative.

While the automated *understanding* of imagery is an extremely challenging long-term goal, the best hope in the near term may be data triage – concede that every image cannot be fully analyzed, but at least identify those images that are most desperately in need of an analyst's attention. This is where change detection can help. While *single* images are indeed difficult to analyze, *pairs* of images can actually be easier. With two images of the same scene, there is a lot of context built right into the data. As the problem goes from finding interesting *things* in images to finding interesting *changes* in image pairs, it becomes more tractable. On the other hand, as Radke *et al.*[1] point out, even the limited aim of detecting changes in imagery is itself extremely challenging.

### 1.1. The case for <u>anomalous</u> change detection

*"Just because everything is different doesn't mean anything has changed."* — Irene Peter[2]

Although there are cases in which an image analyst has a very specific change in mind – *e.g.,* the appearance of camouflaged vehicles or the harvesting of illicit crops – we are concerned with the more open-ended problem of *anomalous* change detection. In this case, there is not an *a priori* target of interest; instead one is looking for changes that are, in some sense, "unusual". Informally, we seek unusual changes because we think they may be indicative of changes that are "meaningful" or "interesting". An unusual change is not necessarily meaningful, but:

---

Emails: (a) jt@lanl.gov, (b) simon.perkins@ultraspectral.com

1. Since unusual changes are by definition rare, the false alarm rate will not be too high.

2. Unlike pervasive changes, which are readily evident by just looking at the imagery, unusual changes point the analyst to spatially small changes that could easily be overlooked.

3. We lack a mathematical characterization of what constitutes a "meaningful" change, but we can be rigorous in our definition of "unusual", and that provides a metric for optimizing algorithms.

The utility of seeking anomalous changes has been emphasized by Schaum and Stocker[3, 4] and by Clifton.[5] In fact, Clifton[5] remarks that sometimes the most interesting change is no change at all – in an area with clear-cut logging, it would be the few trees left standing that would be the unusual changes.

In Ref. [6], we argued that the interesting changes are anomalous changes, and we proposed a framework for anomalous change detection that builds on the machine learning formalism for anomaly detection. We characterize this formulation as distribution-based, rather than difference-based, and in this paper describe several variants of distribution-based anomaly detection.

## 1.2. Notation

We presume that we have two corresponding images. Let $\mathbf{x}_i \in \mathbb{R}^{d_x}$ correspond to the spectra at the $i$th pixel of the first image. Here, $d_x$ is the number of spectral channels in the image: we have $d_x = 1$ for panchromatic imagery, $d_x = 3$ for RGB imagery, and $d_x$ can be a hundred or more for hyperspectral imagery. Let $\mathbf{y}_i \in \mathbb{R}^{d_y}$ be the spectrum at the corresponding pixel in the second image. Some change detection algorithms require that both images have the same number of spectral channels: $d_x = d_y$.

The change detection problem asks: for what pixels $i$ are $\mathbf{x}_i$ and $\mathbf{y}_i$ different; or, more specifically, for what pixels are they "unusually" different.

## 2. DIFFERENCE-BASED CHANGE DETECTION

## 2.1. Simple subtraction

The simplest algorithm for change detection is to subtract the two images, and identify the pixels where the difference has the greatest magnitude. For this to make sense, it is important that the two images be taken with a similar sensor. At the very minimum, we require $d_x = d_y$. And if there is more than one spectral channel, we need to convert the vector difference $\mathbf{e} = \mathbf{y} - \mathbf{x}$ into a scalar measure of anomalousness.[*] A natural choice is the Euclidean distance $\mathbf{e}^T \mathbf{e}$, but a more flexible alternative is the Mahalanobis distance from the centroid of the error vectors. Let $\tilde{\mathbf{e}} = \mathbf{e} - \langle \mathbf{e} \rangle$ be the difference between the difference vector $\mathbf{e}$ and the centroid $\langle \mathbf{e} \rangle$. The angle brackets $\langle \cdot \rangle$ indicate the average over the scene (or possibly over the entire archive). The anomalous measure is given by the Mahalanobis distance $\tilde{\mathbf{e}}^T \langle \tilde{\mathbf{e}} \tilde{\mathbf{e}}^T \rangle^{-1} \tilde{\mathbf{e}}$, where $\langle \tilde{\mathbf{e}} \tilde{\mathbf{e}}^T \rangle$ is a covariance matrix. This is related to the RX anomaly detector,[9] and one way to interpret this choice is that it takes direct differences, but looks for anomalies (outliers) among those differences.

## 2.2. Equalization

Straight differencing can be confounded even by simple miscalibrations in the data. A difference in offset can be taken care of with the subtraction of the centroid ($\tilde{\mathbf{e}} = \mathbf{e} - \langle \mathbf{e} \rangle$ in the above paragraph), but a difference in scale will be problematic. The equalization idea is to bring both data sets to have the same scale and offset.

We apply equalizing functions $H_x(\mathbf{x})$ and $H_y(\mathbf{y})$ to the images separately, and then subtract $H_x(\mathbf{x}) - H_y(\mathbf{y})$. Here $H_x$ and $H_y$ are functions that bring the $\mathbf{x}$ and $\mathbf{y}$ data to a some kind of common foundation. For instance, if $d_x = d_y = 1$, then histogram equalization would bring both datasets to a uniform distribution, and would take out any differences in contrast or calibration.

---

[*]We remark that there is a lot of concern in the literature[7, 8] over finding the threshold for which differences are interesting, but we will take the view that this threshold can always be found manually (or set at a fixed percentile that reflects the resources available to do manual follow-up), and concentrate our efforts on how we measure "anomalousness" in the first place.

Another possibility that might be useful for positive multispectral data would be to normalize each point using $H_x(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$ and similarly for $H_y$. This effectively transforms to a space in which all the data is on the unit sphere, and distances between data samples are spectral angle distances.

For hyperspectral data, it is natural to take $H_x$ and $H_y$ as the functions that respectively whiten the $\mathbf{x}$ and $\mathbf{y}$ data. Used directly, this requires that $d_x = d_y$.[†] This is the idea behind the "covariance equalization" approach.[10] First, subtract the mean values from the data: $\mathbf{x} \leftarrow \mathbf{x} - \langle \mathbf{x} \rangle$ and $\mathbf{y} \leftarrow \mathbf{y} - \langle \mathbf{y} \rangle$. Define the covariance matrices for each image: $X = \langle \mathbf{x}\mathbf{x}^T \rangle$ and $Y = \langle \mathbf{y}\mathbf{y}^T \rangle$. Then the whitened difference is given by

$$\mathbf{e} = Y^{-1/2}\mathbf{y} - X^{-1/2}\mathbf{x} \tag{1}$$

and anomalousness is given by the Mahalanobis distance $\mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e}$.

### 2.2.1. Choosing $R$

The above description is actually a simplified version of what was proposed by Schaum and Stocker.[10] They derive an approximation $\mathbf{y} \approx \hat{\mathbf{y}} = L^*\mathbf{x}$, where $L^*$ is given by the expression $L^* = Y^{1/2}RX^{-1/2}$. This approximation employs an orthonormal matrix $R$ that is not formally specified, but in practice is taken to be $R = I$. In this special case, defining the error of the approximation by $\mathbf{e} = \mathbf{y} - L^*\mathbf{x}$ and defining anomalousness as $\mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e}$ produces the same result as subtracting the whitened images in Eq. (1).

However, this unspecified matrix $R$ provides an opportunity for further optimization. We can interpret this $R$ as a kind of rotation that best "aligns" whitened $X^{-1/2}\mathbf{x}$ with whitened $Y^{-1/2}\mathbf{y}$. Given the identification

$$\mathbf{y} \approx L^*\mathbf{x} = Y^{1/2}RX^{-1/2}\mathbf{x}, \tag{2}$$

or equivalently

$$Y^{-1/2}\mathbf{y} \approx RX^{-1/2}\mathbf{x}, \tag{3}$$

multiply both sides by $\mathbf{x}^T X^{-1/2}$

$$Y^{-1/2}\mathbf{y}\mathbf{x}^T X^{-1/2} \approx RX^{-1/2}\mathbf{x}\mathbf{x}^T X^{-1/2} \tag{4}$$

and then average both sides

$$Y^{-1/2}CX^{-1/2} \approx RX^{-1/2}XX^{-1/2} = R \tag{5}$$

where $C = \langle \mathbf{y}\mathbf{x}^T \rangle$ is a cross-correlation matrix. This suggests an expression for $R$ which is not necessarily orthogonal, but we can "fix" this by using a singular value decomposition of the left hand side. Write

$$Y^{-1/2}CX^{-1/2} = USV^T \tag{6}$$

with $U$ and $V$ orthogonal, and $S$ non-negative and diagonal. Since $S$ corresponds to the "scale" of the matrix, we can factor this out using the simple (but *ad hoc*) approach of setting $S = I$. Specifically, we write $R = UV^T$.

Now this choice of $R$ requires the computation of $C$, which is something that the original covariance equalization took pains to avoid.[‡] But what this provides is a linear anomalous change detector which is potentially better than the standard covariance equalization. Furthermore, when $d_x \neq d_y$, this method prescribes a rectangular $R$ that enables the covariance equalization approach to address the $d_x \neq d_y$ case. For the panchromatic $d_x = d_y = 1$ case, $R$ is restricted to scalar values $\pm 1$. The advantage of using this $R$ instead of $R = I$ is illustrated in Fig. 2(a).

---

[†]One can imagine variants for dealing with $d_x \neq d_y$, such as taking the first $d \leq \min(d_x, d_y)$ principal components from each image before whitening, but since the approach is based on subtracting imagery, it is probably a good idea to use the same modality for both the $\mathbf{x}$ and $\mathbf{y}$ images, and in practice, this usually implies $d_x = d_y$ to begin with.

[‡]Because the computation of the cross-covariance $C$, unlike that of the covariance matrices $X$ and $Y$, requires co-registration of the images, avoiding this computation leads to a linear map $L^*$ that is robust to misregistration.[10] On the other hand, the actual difference operation in Eq. (1) does require good co-registration; and the estimation of $\langle \mathbf{e}\mathbf{e}^T \rangle = 2I - Y^{-1/2}CX^{-1/2} - X^{-1/2}C^T Y^{-1/2}$ is clearly related to the estimation of $C$.

## 2.3. Regression

Where equalization tries to adapt both images to a common foundation, regression attempt to directly fit a function that transforms one image into the other. Informally speaking, this function is the "correction" that accounts for the different imaging conditions in the second image *vis-a-vis* the first. Clifton[5] used a neural network to learn a nonlinear relationship between $\mathbf{x}$ and $\mathbf{y}$ and then argued that "substantial differences between the expected and actual values represent an unusual change." The idea is to fit a function $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ such that $\mathbf{y} \approx \hat{\mathbf{y}} = f(\mathbf{x})$ is a good approximation, and then to compute differences

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - f(\mathbf{x}). \tag{7}$$

These differences can then be assessed with the usual Mahalanobis approach, by looking at the scalar values of $\mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e}$.

The linear variant of this, as applied in hyperspectral remote sensing, goes by the name *chronochrome*.[4] Here, the idea is to fit a linear map $L$ such that

$$\mathbf{y} \approx \hat{\mathbf{y}} = L\mathbf{x}. \tag{8}$$

Here, $L$ is fit using least-squares. Defining an error matrix

$$E = \left\langle (\mathbf{y} - L\mathbf{x})(\mathbf{y} - L\mathbf{x})^T \right\rangle, \tag{9}$$

one can show that the trace of $E$ is minimized when $L = CX^{-1}$. The essence of the change detection algorithm is to find changes between $\mathbf{y}$ and $\hat{\mathbf{y}} = L\mathbf{x}$. Write

$$\mathbf{e} = \mathbf{y} - CX^{-1}\mathbf{x} \tag{10}$$

and compute the covariance $\left\langle \mathbf{e}\mathbf{e}^T \right\rangle = Y - CX^{-1}C^T$. The largest changes $\mathbf{e}$, in the sense of Mahalanobis distance, are given by the largest values of $\mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle \mathbf{e}$, or

$$(\mathbf{y} - CX^{-1}\mathbf{x})^T (Y - CX^{-1}C^T)^{-1}(\mathbf{y} - CX^{-1}\mathbf{x}). \tag{11}$$

The comparison of this quantity to an appropriate threshold defines the chronochrome anomaly detector. The boundary that separates anomalous changes from normal is an elliptical cylinder centered on the axis defined by $\mathbf{y} = CX^{-1}\mathbf{x}$.

An idiosyncrasy of the chronochrome is that it is asymmetric with respect to $\mathbf{x}$ and $\mathbf{y}$. If one instead looks for a linear map such that $\mathbf{x} \approx \hat{\mathbf{x}} = L'\mathbf{y}$, and minimizes $|\mathbf{x} - L'\mathbf{y}|^2$, then one obtains a different chronochrome detector: namely,

$$(\mathbf{x} - C^T Y^{-1}\mathbf{y})^T (X - C^T Y^{-1}C)^{-1}(\mathbf{x} - C^T Y^{-1}\mathbf{y}), \tag{12}$$

which leads to a different cylinder, whose axis is specified by $\mathbf{x} = C^T Y^{-1}\mathbf{y}$.

## 3. DISTRIBUTION-BASED CHANGE DETECTION

Unlike the approaches in the previous section, the distribution-based algorithms are not built around a big subtraction step in the middle. Here, we treat $\mathbf{x}$ and $\mathbf{y}$ values as if drawn from a parent distribution $P(\mathbf{x}, \mathbf{y})$, and attempt to infer anomalous changes based on properties of that distribution. In keeping with a central tenet of machine learning philosophy,[11] we infer these properties without directly estimating $P(\mathbf{x}, \mathbf{y})$ itself. In particular, by recasting the anomalous change detection problem in terms of a ratio of distributions, we can employ proven machine learning algorithms to identify optimal boundaries in the $(\mathbf{x}, \mathbf{y})$ space.

## 3.1. Straight anomaly detection

Application of standard anomaly detection in the combined $(\mathbf{x}, \mathbf{y})$ space leads to contours of constant $P(\mathbf{x}, \mathbf{y})$; these are optimal in the sense of enclosing the maximum fraction of the normal data with the minimum volume. Fig. 5(a,e) illustrate this anomaly detector for two elliptically-contoured distributions. Since the boundaries defining anomalies are defined by the contours of the distribution, we have, in these two cases, that those boundaries are elliptical.

Anomaly (aka novelty) detection has been developed in the machine learning community,[12–15] for applications from handwriting characterization[13] to malicious executable detection.[16]

By writing $P(\mathbf{x}, \mathbf{y}) \propto P(\mathbf{x}, \mathbf{y})/U(\mathbf{x}, \mathbf{y})$ where $U(\mathbf{x}, \mathbf{y})$ is a constant, or uniform, distribution, we express the contours we seek in terms of a ratio of distributions; by sampling points from these two distributions: one set from $P(\mathbf{x}, \mathbf{y})$ and another from $(U\mathbf{x}, \mathbf{y})$, we set up a classification problem whose optimal solution is the contours we seek. The samples from $P(\mathbf{x}, \mathbf{y})$ are just the data from the two images, and the samples from $U(\mathbf{x}, \mathbf{y})$ are just points sampled uniformly over a subset of $\mathbb{R}^{d_x + d_y}$ that encloses the support of $P(\mathbf{x}, \mathbf{y})$.

The casting of anomaly detection as a binary classification problem, with one class provided by the available (and presumed normal) data, and the other class defined in terms of a uniformlly random background distribution, was discussed and illustrated by Hastie *et al.*,[17] specifically see Chapter 14.2.4 and especially Fig 14.3, and put on more formal mathematical foundation by Steinwart *et al.*.[14]

As has been emphasized elsewhere,[14, 18] these contours depend on an underlying background measure. Traditionally, the standard Lebesgue measure is used, but it is difficult to justify this choice in any formal way.

We have not seen this straight anomaly detection approach advocated for the change detection problem, and that's because straight anomaly detection does not consider the *change* in going from $\mathbf{x}$ to $\mathbf{y}$; it finds all anomalies, but we want to restrict our attention to the anomalous changes.

## 3.2. A generalized chronochrome

If we want to know whether $\mathbf{y}$ is anomalous, *conditioned* on $\mathbf{x}$, then we can use the straight anomaly detection approach described above, but only in the $\mathbf{y}$ component. In particular, we are concerned with contours of the conditional distribution $P(\mathbf{y}|\mathbf{x}) = P(\mathbf{x}, \mathbf{y})/P_x(\mathbf{x})$. Here, $P_x(\mathbf{x})$ is the marginal distribution, given by

$$P_x(\mathbf{x}) = \int P(\mathbf{x}, \mathbf{y}) \, d\mathbf{y}. \tag{13}$$

Note that we can write $P(\mathbf{y}|\mathbf{x}) \propto P(\mathbf{x}, \mathbf{y})/\left[P_x(\mathbf{x})U(\mathbf{y})\right]$, where $U(\mathbf{y})$ is a constant, and corresponds to a uniform distribution over $\mathbf{y}$. Again, we are seeking contours of the ratio of two probability distributions: $P(\mathbf{x}, \mathbf{y})$ and $P_x(\mathbf{x})U(\mathbf{y})$, and again, this formulation permits us to employ the formalism of binary classification. The two classes correspond to the distributions $P(\mathbf{x}, \mathbf{y})$ and $P_x(\mathbf{x})U(\mathbf{y})$. Samples from $P(\mathbf{x}, \mathbf{y})$ are just given by the data; samples from $P_x(\mathbf{x})U(\mathbf{y})$ can be generated from points $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is taken from random data points, and $\mathbf{y}$ is chosen uniformly over an interval of $\mathbf{y}$ that includes the support of $P(\mathbf{x}, \mathbf{y})$. With this resampling scheme, we never actually estimate either $P(\mathbf{x}, \mathbf{y})$ or $P_x(\mathbf{x})$.

We remark that when the data are Gaussian, then this approach leads to the linear chronochrome. Thus, it provides a natural generalization of the chronochrome for non-Gaussian data. And like the linear chronochrome, there is an asymmetry in $\mathbf{x}$ and $\mathbf{y}$. We could just as well seek contours of $P(\mathbf{x}|\mathbf{y})$, and this would tell us when $\mathbf{x}$ was anomalous, conditioned on $\mathbf{y}$.

We further remark that this asymmetry has an arrow-of-time flavor. If the $\mathbf{x}$ image were taken before the $\mathbf{y}$ image, it seems more natural to consider $P(\mathbf{y}|\mathbf{x})$, which – informally speaking – emphasizes finding anomalous $\mathbf{y}$'s conditioned on normal $\mathbf{x}$'s.
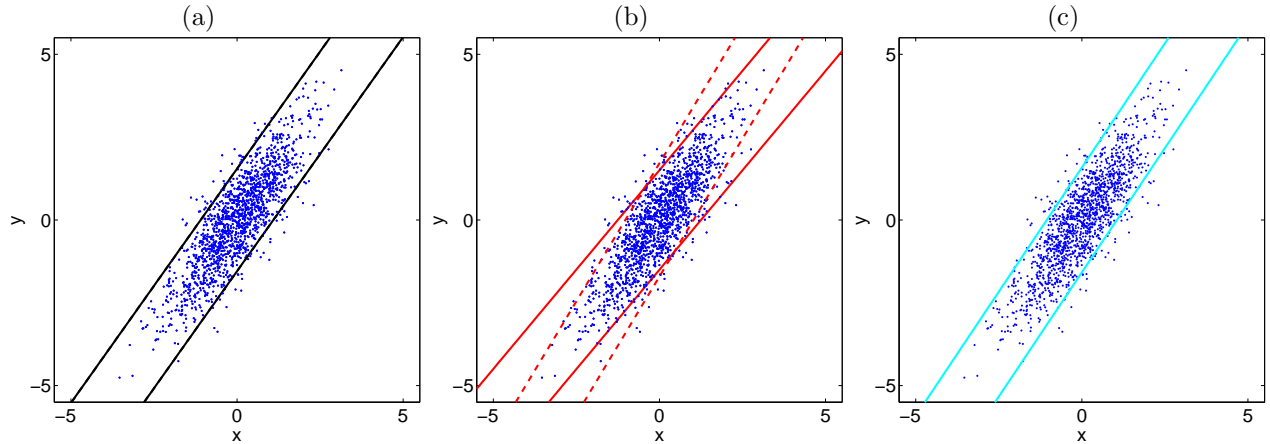
**Figure 1.** Detection boundaries are shown for different anomalous change detectors, applied to a Gaussian distribution with $d_x = d_y = 1$. The data are shown as $N = 500$ points sampled from the normal distribution $P(x, y)$, parameterized by the covariance matrix in Eq. (15) with $X = 1$, $Y = 2$, and $C = 1.2$. The boundary thresholds are chosen so that the same fraction (95%) of data points are within the normal regime. **(a)** Covariance equalization, based on Eq. (1). **(b)** Chronochrome, with the solid line corresponding to $\mathbf{y}$ fit as a function of $\mathbf{x}$ (Eq. (11)), and the dashed line to $\mathbf{x}$ fit as a function of $\mathbf{y}$ (Eq. (12)). **(c)** The minimum-width linear parallel boundaries.

## 3.3. Anomalous change detection with a generalized "achronochrome"

A natural way to symmetrize the generalized chronochrome is to consider the ratio of distributions given by

$$\frac{P(\mathbf{x}, \mathbf{y})}{P_x(\mathbf{x})P_y(\mathbf{y})} \tag{14}$$

Here, $P_x(\mathbf{x})P_y(\mathbf{y})$ is the distribution that would be exhibited by the data if $\mathbf{x}$ and $\mathbf{y}$ were independent. By using this as the background, we obtain anomalies whose $(\mathbf{x}, \mathbf{y})$ dependency is unusual compared to the dependency encoded in $P(\mathbf{x}, \mathbf{y})$.

It is straightforward to turn this ratio into a resampling algorithm, and we have done so previously.[6] It is (once again) binary classification: one set of samples comes directly from the data, specifically $(\mathbf{x}, \mathbf{y})$ pairs with $\mathbf{x}$ and $\mathbf{y}$ taken from corresponding pixels in the two images; the other set is resampled from the data by taking $\mathbf{x}$ from one pixel and $\mathbf{y}$ from a different pixel.

This eliminates the arbitrary choice of a uniform background entirely. Points $(\mathbf{x}, \mathbf{y})$ are considered an anomalous change pair only if $P(\mathbf{x}, \mathbf{y})$ is small and $P_x(\mathbf{x})$ and $P_y(\mathbf{y})$ are relatively large. That is, $\mathbf{x}$ and $\mathbf{y}$ are individually normal, but their *relationship* is unusual.

What is happening is that "rules" are learned from the data which characterize the relationship $\mathbf{x} \leftrightarrow \mathbf{y}$ because that occur over a large number of pixels. Those pixels for which the $\mathbf{x} \leftrightarrow \mathbf{y}$ relationship is inconsistent with this rule are identified as anomalous changes. The framework is quite general – the two images need not have the same (or even the same number of) spectral bands. Furthermore, there is complete symmetry in the $\mathbf{x} \leftrightarrow \mathbf{y}$. This lack of a sense of temporal direction suggests the name *a*chronochrome.

## 4. EXAMPLES

### 4.1. Gaussian Distribution

To make the anomalous change detection framework a little more concrete, and to connect the approach with some existing methods that have proved successful for hyperspectral data, we will consider the case in which the measured data are drawn from Gaussian distributions. We emphasize that the distribution-based approaches do not depend on an assumption of Gaussianity.
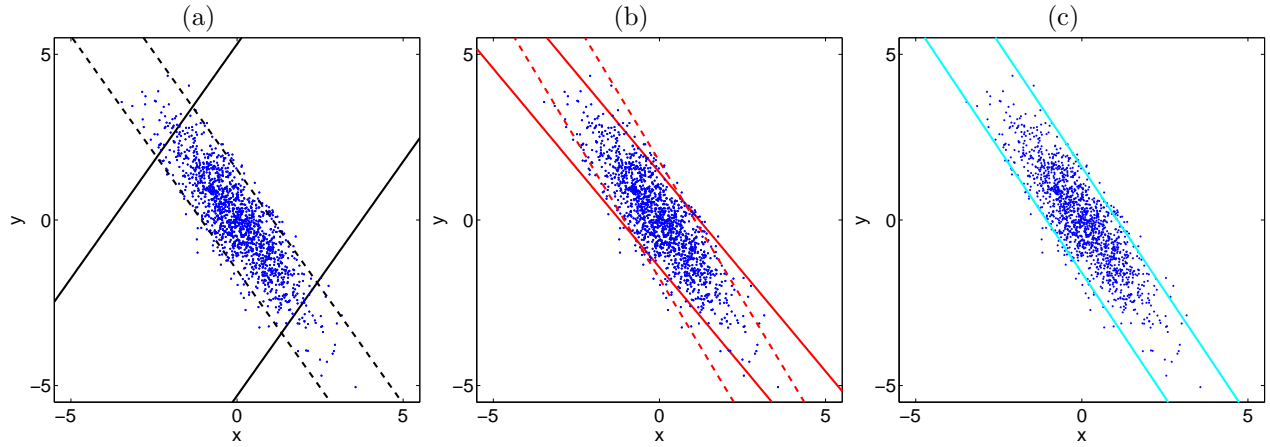
**Figure 2.** Same as Fig. 1, but for negative correlation $C = -1.2$. **(a)** In this case, standard covariance equalization (solid lines) produces a very wide pair of parallel lines, whereas the $R$ corrected approach (dashed lines) more nearly "follow" the flow of the data. **(b)** Both variants of the chronochrome follow the negative correlation. **(c)** The minimum-width linear parallel boundaries.
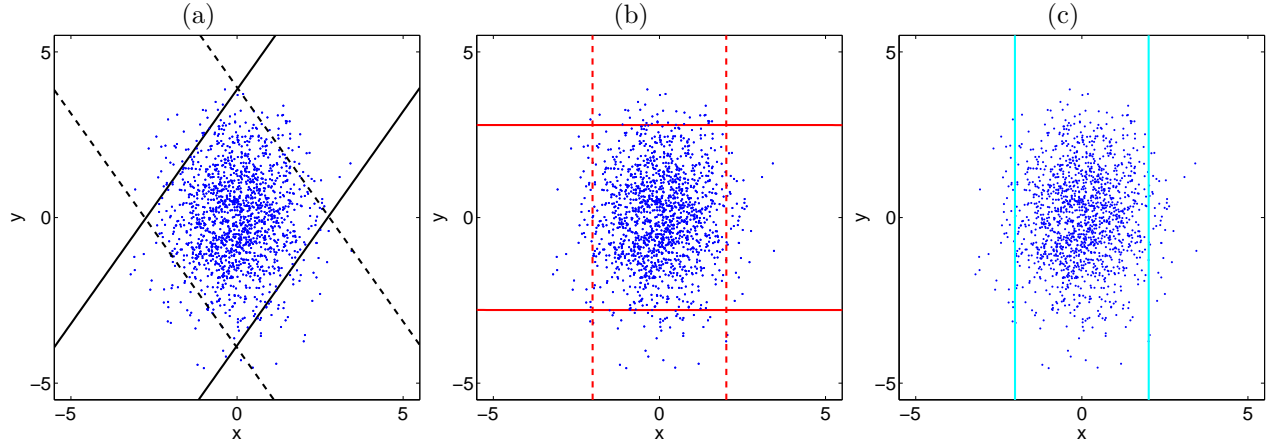


**Figure 3.** Same as Fig. 1, with no correlation between the images: $C = 0$.

We presume that the points $(\mathbf{x}, \mathbf{y})$ are drawn from a $d_x + d_y$ dimensional Gaussian distribution with mean zero and covariance given by the matrix

$$\left\langle \left[ \begin{array}{c} \mathbf{x} \\ \mathbf{y} \end{array} \right] [\mathbf{x}^T \, \mathbf{y}^T] \right\rangle = \left[ \begin{array}{cc} X & C^T \\ C & Y \end{array} \right] \tag{15}$$

where superscript $T$ indicates transpose, and

$$X = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle, \tag{16}$$
$$Y = \left\langle \mathbf{y}\mathbf{y}^T \right\rangle, \tag{17}$$
$$C = \left\langle \mathbf{y}\mathbf{x}^T \right\rangle. \tag{18}$$

The normal points are drawn from the distribution $P(\mathbf{x}, \mathbf{y})$, which is Gaussian with mean zero and covariance given by the matrix in Eq. (15). The goal is to identify boundaries in the combined $(\mathbf{x}, \mathbf{y})$ space (that is, in $\mathbb{R}^{d_x + d_y}$) which enclose most of these normal points, but outside of which anomalous changes will be found. The special case $d_x = d_y = 1$ leads to two-dimensional plots, and these are shown in Figs. 1-3.

### 4.1.1. Difference-based anomaly detectors

For all three of these plots, the (a) panel shows the covariance equalization boundary. The solid curve is the based on the standard whitened difference in Eq. (1), and the dashed curve uses the variant with the $R$ correction. When the correlation is positive, as seen in Fig. 1(a), $R = 1$, and both variants give the same result. But when the correlation is negative, $R = -1$ and Fig. 2(a) shows that the $R$-corrected approach gives much narrower boundaries than the standard. In Fig. 3(a), the dashed curve is drawn for $R = -1$, but since this case corresponds to a cross-covariance of zero, and there is ambiguity in assigning $R = \pm 1$.

The (b) panel in these plots shows the chronochrome boundaries for the two cases in Eq. (11) (solid) and Eq. (12) (dashed).

Panel (c) is the minimum-distance parallel boundaries; this is in some sense the optimal linear solution. In general, this minimum-distance solution differs from either the covariance equalization or the chronochrome. (In Ref. [6] we incorrectly stated that the covariance equalization algorithm gave the minimum-distance solution; this is true when $X = Y$, but it is not generally the case.)

### 4.1.2. Distribution-based anomaly detectors

We describe how the proposed anomalous change detection framework would be applied to the linear Gaussian model. Here, $P(\mathbf{x}, \mathbf{y})$ is Gaussian with covariance given by Eq. (15), and $P(\mathbf{x})$ and $P(\mathbf{y})$ are Gaussian with covariances given by $X$ and $Y$, respectively. In this framework, anomalous changes are given by level curves of Eq. (14); up to translation and scale factors, the negative logarithm of Eq. (14) is the quadratic function

$$\begin{bmatrix} \mathbf{x}^T & \mathbf{y}^T \end{bmatrix} K \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{19}$$

where the matrix $K$ is given by

$$K = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}^{-1} - \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix}^{-1} \tag{20}$$

Fig. 5(h) illustrates what these boundaries would be for the simple case of single-band images ($d_x = d_y = 1$) with $X = 1$, $Y = 2$, and $C = 1.2$. Note that $K$ is not a positive-definite matrix; so the contours in Eq. (19) are hyperbolic, not elliptical.

### 4.2. EC Distribution

Following Manolakis *et al.*,[19–21] we consider an elliptically-contoured (EC) multivariate $t$ distribution, with $\nu$ degrees of freedom. Since the $\nu \to \infty$ limit leads to a Gaussian distribution, we can think of this EC distribution as a generalization of the Gaussian. Fig. 5(a,b,c,d) compares the boundaries of an EC to those of a Gaussian distribution. The generalized chronochrome is seen to give curved lines in this case, but qualitatively it the boundaries are similar.

## 5. CONCLUSION

We described a taxonomy of anomalous change detection algorithms, and drew a distinction between difference-based and distribution-based approaches. The new distribution-based framework for anomalous change detection permits us both to derive standard approaches, such as the chronochrome, to generalize them to non-gaussian situations, and to introduce new algorithms.

Our exposition emphasized spectral ("chromatic") over spatial issues, but we recognize the importance in imagery of taking into account not only coregistration and ground sample distances issues, but spatial correlations, textures, and proximities as well.

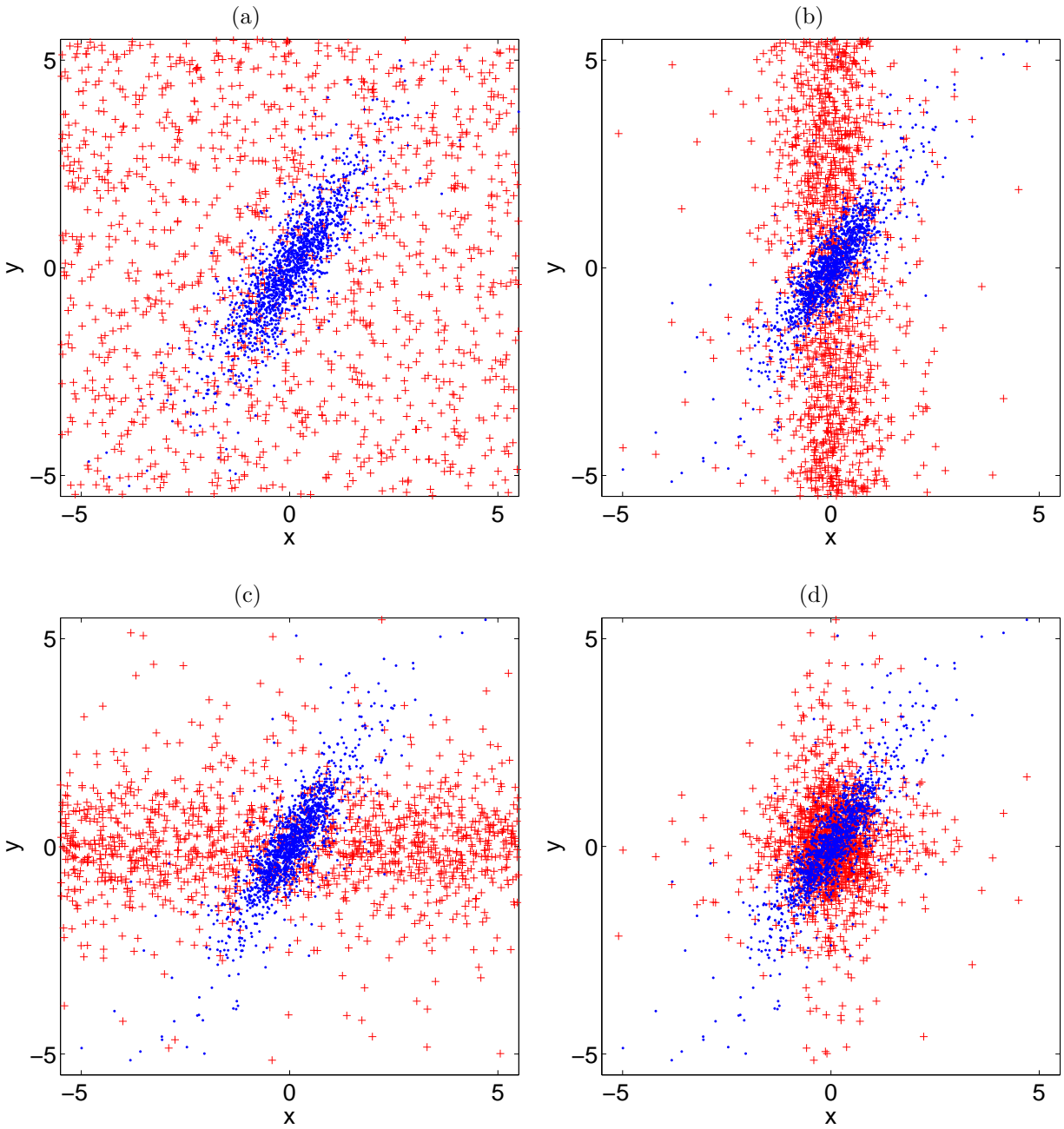The work we have described is preliminary. Further efforts are being pursued along the following lines:

**Figure 4.** Resampling: by resampling from the data or from a uniform distribution in either or both of the **x** or **y** axes, we recast the anomalous change detection problem to one of binary classification. The data (dots) are to be classified as separate from the background (crosses). **(a)** Straight anomaly detection; **(b)** Chronochrome (**y** fit as a function of **x**); **(c)** Chronochrome (**x** fit as a function of **y**); and **(d)** "Achronochrome" detector described in Section 3.3.
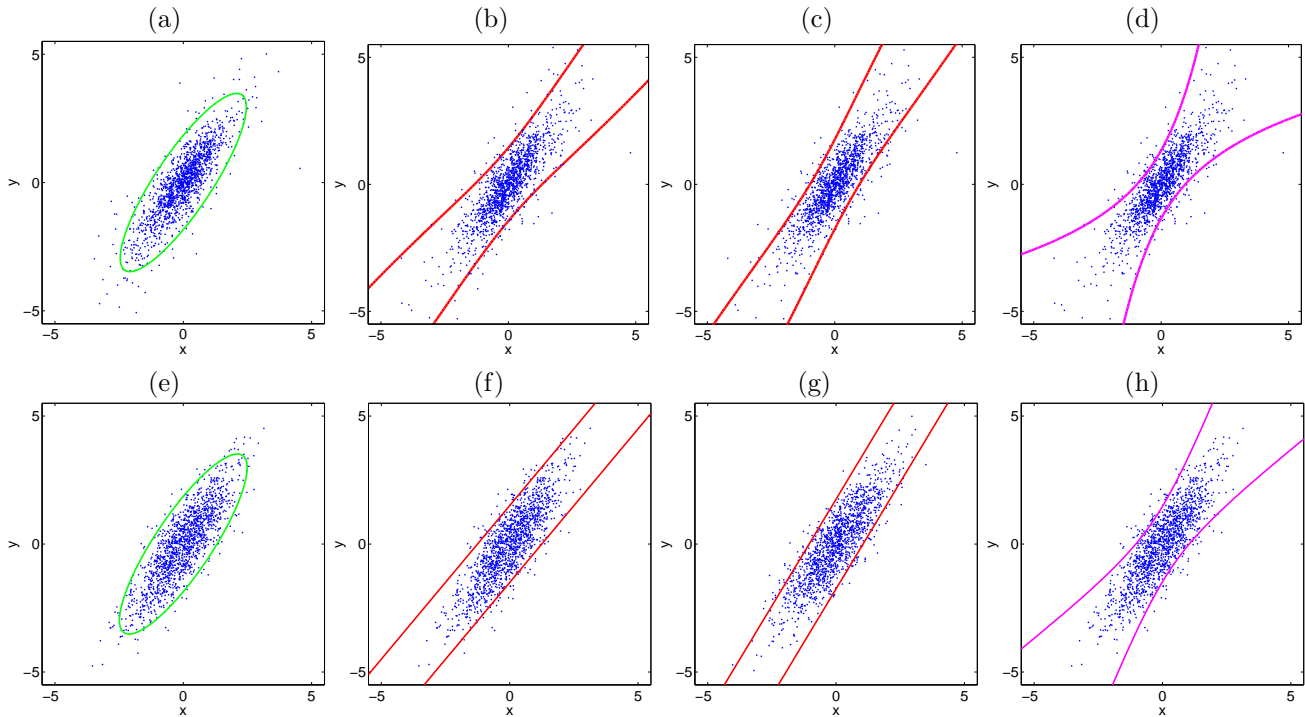
**Figure 5.** Illustrating boundaries for distribution-based anomalous change detectors. **(a,b,c,d)** Distribution is an EC multivariate $t$ distribution with $\nu = 5$ degrees of freedom; **(e,f,g,h)** Distribution is gaussian. In both cases, the covariance matrix uses $X = 1$, $Y = 2$, $C = 1.2$ (same as Fig. 1). The data are shown as $N = 500$ points sampled from the distribution $P(x, y)$. In all cases, the boundaries are drawn so that they enclose 95% of the data **(a,e)** Straight anomaly detection; **(b,f)** Chronochrome (**y** fit as a function of **x**); **(c,g)** Chronochrome (**x** fit as a function of **y**); and **(d,h)** "Achronochrome" detector described in Section 3.3.

- Any change-detection scheme, ours included, benefits when the two images are better matched to begin with. We have employed genetic programming[22] to learn image processing pipelines that lead to good classifiers. It is possible that a similar scheme can be used to optimize the choice of these spatial operators for pre-processing, to create better-matched images before the spectral anomaly detection step is invoked.

- Pixels are highly correlated, and associated with an arbitrary grid associated not with the scene, but with the geometry of the focal plane on the camera. Triangulating the image pairs into a common mesh of "trixels," which can be aggregated into larger polygons, provides a characterization that has more of an object-based flavor.[23]

- One approach for dealing with small misregistrations in conventional change detection is to compare differences between a pixel in one image with each of the pixels in a small neighborhood of the corresponding pixel in the other image. The appropriate generalization of this idea to the distribution-based approaches would employ multiple-instance learning.[24]

- Kwon *et al.* has advocated the use of kernels for anomaly detection[25]; the extension to anomalous change detection should be relatively straightforward, and merits further investigation.

- Sometimes there are interesting anomalies that are not that rare. Rather than simply rank all pixels on a scalar value of anomalousness, it may be productive to cluster the anomalies.[26]

## REFERENCES

1. R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Processing* **14**, pp. 294–307, 2005.

 2. Quoted on wikipedia: `http://en.wikiquote.org/wiki/Change`.

 3. A. Schaum and A. Stocker, "Subclutter target detection using sequences of thermal infrared multispectral imagery," *Proc. SPIE* **3071**, pp. 12–22, 1997.

 4. A. Schaum and A. Stocker, "Long-interval chronochrome target detection," *Proc. 1997 International Symposium on Spectral Sensing Research* , 1998.

 5. C. Clifton, "Change detection in overhead imagery using neural networks," *Applied Intelligence* **18**, pp. 215–234, 2003.

 6. J. Theiler and S. Perkins, "Proposed framework for anomalous change detection," *ICML Workshop on Machine Learning Algorithms for Surveillance and Event Detection* , pp. 7–14, 2006.

 7. L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geoscience and Remote Sensing* **38**, pp. 1171–1182, 2000.

 8. F. Melgani and Y. Bazi, "Markovian fusion approach to robust unsupervised change detection in remotely sensed imagery," *IEEE Geoscience and Remote Sensing Letters* **3**, pp. 457–461, 2006.

 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* **38**, pp. 1760–1770, 1990.

10. A. Schaum and A. Stocker, "Hyperspectral change detection and supervised matched filtering based on covariance equalization," *Proc. SPIE* **5425**, pp. 77–90, 2004.

11. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 2nd ed., 1999.

12. S. Ben-David and M. Lindenbaum, "Learning distributions by their density levels: A paradigm for learning without a teacher," *J. Computer and System Sciences* **55**, pp. 171–182, 1997.

13. B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation* **13**, pp. 1443–1471, 2001.

14. I. Steinwart, D. Hush, and C. Scovel, "Density level detection is classification," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, eds., pp. 1337–1344, MIT Press, Cambridge, MA, 2005.

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

16. D. M. Cai, J. Theiler, and M. Gokhale, "Detecting a malicious executable without prior knowledge of its patterns," *Proc. SPIE* **5812**, pp. 1–12, 2005.

17. T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, 2001.

18. J. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," *Proc. SPIE* **5093**, pp. 230–240, 2003.

19. D. Manolakis, D. Marden, J. Kerekes, and G. Shaw, "On the statistics of hyperspectral imaging data," *Proc. SPIE* **4381**, pp. 308–316, 2001.

20. D. B. Marden and D. Manolakis, "Modeling hyperspectral imaging data," *Proc. SPIE* **5093**, pp. 253–262, 2003.

21. D. B. Marden and D. Manolakis, "Using elliptically contoured distributions to model hyperspectral imaging data and generate statistically similar synthetic data," *Proc. SPIE* **5425**, pp. 558–572, 2004.

22. N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J. J. Bloch, R. B. Porter, M. Galassi, and A. C. Young, "Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction," *IEEE Trans. Geosci. and Remote Sens.* **40**, pp. 393–404, 2002.

23. L. Prasad and A. Skourikhine, "Vectorized image segmentation via trixel agglomeration," *Pattern Recognition* **39**, pp. 501–514, 2006.

24. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Prez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence* **89**, pp. 31–71, 1997.

25. H. Kwon and N. M. Nasrabadi, "Kernel RX: a new nonlinear anomaly detector," *Proc. SPIE* **5806**, pp. 35–46, 2005.

26. D. Pelleg and A. W. Moore, "Active learning for anomaly and rare-category detection," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, eds., pp. 1073–1080, MIT Press, Cambridge, MA, 2005.