Total Least Squares for Anomalous Change Detection

James Theiler and Anna M. Matsekh

Space and Remote Sensing Sciences, Los Alamos National Laboratory, Los Alamos, NM, USA

ABSTRACT

A family of subtraction-based anomalous change detection algorithms is derived from a total least squares (TLSQ) framework. This provides an alternative to the well-known chronochrome algorithm, which is derived from ordinary least squares. In both cases, the most anomalous changes are identified with the pixels that exhibit the largest residuals with respect to the regression of the two images against each other. The family of TLSQ-based anomalous change detectors is shown to be equivalent to the subspace RX formulation for straight anomaly detection, but applied to the stacked space. However, this family is not invariant to linear coordinate transforms. On the other hand, *whitened* TLSQ is coordinate invariant, and special cases of it are equivalent to canonical correlation analysis and optimized covariance equalization. What whitened TLSQ offers is a generalization of these algorithms with the potential for better performance.

Keywords: total least squares, change detection, chronochrome, covariance equalization, anomaly detection

1. SAMUEL BECKETT'S INTRODUCTION

Scene one. Images two. Call them x and y. Taken at different times. On different days. Possibly even with different cameras. Different scene illumination, sensor calibration, atmospheric conditions. Imperfectly registered. Everything is different. But there is something else. Something has changed. Something interesting, rare, unusual. Hidden in plain sight. Anomalous change among the pervasive differences.

2. TECHNICAL INTRODUCTION

The main challenge for the anomalous change detection (ACD) problem is to distinguish the anomalous changes (which are rare) from the pervasive differences (which occur throughout the scene). Since virtually all of the data corresponds to pervasive differences, a model that is fit to the data will be a model of pervasive differences. The pixels that deviate from this model are then flagged as candidates for anomalous changes. ACD based on a linear fit to the data was introduced by Schaum and Stocker[1,2] and alliteratively denominated the "chronochrome". An extension to nonlinear fits using neural networks was described by Clifton.[3] Further variations include covariance equalization,[4,5] and multivariate alteration detection,[6] which is based on canonical components analysis.[7] The notion of distinguishing pervasive differences from anomalous changes led to a more formal machine learning framework[8,9] which recast the problem as one of binary classification. A recent survey of the practical issues that arise in the ACD problem is given by Eismann *et al.*[10]

The presentation here in terms of total least squares[11,12] extends (and symmetrizes) the chronochrome, but also connects to covariance equalization and multivariate alteration detection. All of these belong to the family of quadratic covariance-based ACD algorithms.[13]

Let $\mathbf{x} \in \mathbb{R}^{d_x}$ be the spectrum of a pixel in the x-image, and $\mathbf{y} \in \mathbb{R}^{d_y}$ the spectrum of the corresponding pixel in the y-image. We will assume, without loss of generality, that means have been subtracted from $\mathbf{x}$ and $\mathbf{y}$. Thus, $\langle \mathbf{x} \rangle = 0$ and $\langle \mathbf{y} \rangle = 0$, where the angle brackets indicate an average over pixels (*e.g.,* over all the pixels in an image). And we will write the following covariance and cross-covariances:

$$X = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle, \tag{1}$$

$$Y = \left\langle \mathbf{y}\mathbf{y}^T \right\rangle, \tag{2}$$

$$C = \left\langle \mathbf{y}\mathbf{x}^T \right\rangle, \tag{3}$$

Email: {jt,matsekh}@lanl.gov

1. Identify two matrices, B_x and B_y, with k columns and d_x and d_y rows, respectively; how these matrices are chosen varies from method to method, but most employ information about the variance and covariances of the data.
2. At each pixel pair $(\mathbf{x}, \mathbf{y})$, compute transformed pixel values: $\mathbf{x}' = B_x^T \mathbf{x}$ and $\mathbf{y}' = B_y^T \mathbf{y}$. Note that both $\mathbf{x}'$ and $\mathbf{y}'$ are k-dimensional.
3. Subtract the transformed pixels: $\mathbf{e} = \mathbf{y}' - \mathbf{x}'$
4. Compute the $k \times k$ covariance matrix of the difference vectors: $\langle \mathbf{e}\mathbf{e}^T \rangle$.
5. Associate the "anomalousness" of a vector-valued pixel-pair difference $\mathbf{e}$ with its Mahalanobis distance to the origin: $\mathcal{A}(\mathbf{e}) = \mathbf{e}^T \langle \mathbf{e}\mathbf{e}^T \rangle^{-1} \mathbf{e}$.

Figure 1. Framework for subtraction-based change detection

where the superscript T indicates transpose. It is useful to write

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \in \mathbb{R}^d, \tag{4}$$

with $d = d_x + d_y$, as the pixel in the "stacked" image. Again, we take $\langle \mathbf{z} \rangle = 0$, and we can write the stacked covariance matrix*

$$Z = \langle \mathbf{z}\mathbf{z}^T \rangle = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}. \tag{5}$$

It is also useful to define the data matrix

$$D_x = [\mathbf{x}_1 \; \mathbf{x}_2 \; \cdots \; \mathbf{x}_n], \tag{6}$$

where $\mathbf{x}_i$ corresponds to the ith pixel in the image, and n is the total number of pixels in the dataset. We can similarly define D_y and D_z, and this allows an alternative set of expressions for covariances and cross-covariances:

$$X = (1/n) D_x D_x^T, \tag{7}$$
$$Y = (1/n) D_y D_y^T, \tag{8}$$
$$C = (1/n) D_y D_x^T, \tag{9}$$
$$Z = (1/n) D_z D_z^T. \tag{10}$$

3. SUBTRACTION-BASED CHANGE DETECTION

The simplest way to detect differences between two images is to subtract them. Where the images are identical, the difference will be zero. Larger differences are naturally interpreted as more important changes. For a variety of reasons, this direct approach doesn't always work very well. When $d_x \neq d_y$, it's not even possible. Nonetheless, most change detections algorithms are based on this subtraction principle. In fact, a number of them (including chronochrome, covariance equalization, multivariate alteration detection, and as we will show, total least squares) employ the framework shown in Fig. 1. In this framework, $\mathbf{x}$ and $\mathbf{y}$ are first transformed, and the transformed values are subtracted. The vector valued difference $\mathbf{e} = \mathbf{y}' - \mathbf{x}'$ is converted to a scalar by computing Mahalanobis distance to the origin. This framework allows us to express the anomalousness of a pixel pair $(\mathbf{x}, \mathbf{y})$ as an explicit function of B_x, B_y, and the covariances X, Y, and C. In particular,

$$\mathcal{A}(\mathbf{x}, \mathbf{y}) = (B_y^T \mathbf{y} - B_x^T \mathbf{x})^T (B_y^T Y B_y - B_y^T C B_x - B_x^T C^T B_y + B_x^T X B_x)^{-1} (B_y^T \mathbf{y} - B_x^T \mathbf{x}). \tag{11}$$

*We will henceforth assume that the covariance matrices X, Y, and Z are of full rank and have well-defined inverses. If that were not the case, it would mean that we had some strictly redundant bands (*i.e.,* bands that were linear combinations of other bands), which we could then just get rid of, and achieve our assumed full-rank state.

This expression is quadratic in $\mathbf{x}$ and $\mathbf{y}$, and can be expressed in terms of a $d \times d$ matrix Q, so that

$$\mathcal{A}(\mathbf{x}, \mathbf{y}) = \left[\begin{array}{cc} \mathbf{x}^T & \mathbf{y}^T \end{array} \right] \; Q \; \left[\begin{array}{c} \mathbf{x} \\ \mathbf{y} \end{array} \right] = \mathbf{z}^T Q \mathbf{z}, \tag{12}$$

where $\mathbf{z}$ is the stacked pixel defined in Eq. (4), and Q is a positive semi-definite matrix:

$$Q = \left[\begin{array}{c} -B_x \\ B_y \end{array} \right] \left[B_y^T Y B_y - B_y^T C B_x - B_x^T C^T B_y + B_x^T X B_x \right]^{-1} \left[\begin{array}{cc} -B_x^T & B_y^T \end{array} \right]. \tag{13}$$

If we write

$$B = \left[\begin{array}{c} -B_x \\ B_y \end{array} \right], \tag{14}$$

then the residuals $\mathbf{e} = \mathbf{y}' - \mathbf{x}' = B_y^T \mathbf{y} - B_x^T \mathbf{x}$ can be simply written $\mathbf{e} = B^T \mathbf{z}$. And

$$Q = B \left(B^T \left[\begin{array}{cc} X & C^T \\ C & Y \end{array} \right] B \right)^{-1} B^T = B(B^T Z B)^{-1} B^T. \tag{15}$$

Since B is a $d \times k$ matrix, $B^T Z B$ is $k \times k$, which means that Q is at most rank k. If B is of lower rank than k, that will lead to the matrix $B^T Z B$ having less than full rank, and therefore not being strictly invertible. The operational solution is to take the *pseudoinverse*.[14, 15] Specifically, if $k' \leq k$ is the rank of B, then we can write $B = B_o R^T$ where B_o is a $d \times k'$ matrix of rank k' and R is a $k \times k'$ matrix satisfying $R^T R = I$. In place of $(B^T Z B)^{-1}$, we use $R(B_o^T Z B_o)^{-1} R^T$. This leads to an expression for the matrix

$$Q = B \, \texttt{pseudoinverse}(B^T Z B) \, B^T = B_o R^T R (B_o^T Z B_o)^{-1} R^T R B_o = B_o (B_o^T Z B_o)^{-1} B_o^T. \tag{16}$$

As Eq. (15) indicates, the subtraction-based ACD algorithm is defined by the matrix B. But invertible linear transforms of B give equivalent detectors. If G is an invertible $k \times k$ matrix, then the transform $B \to BG^T$ leaves Q invariant. That is,

$$\begin{aligned} Q_G &= BG^T \left((BG^T)^T Z (BG^T) \right)^{-1} (BG^T)^T = BG^T (GB^T Z B G^T)^{-1} G B^T = BG^T G^{-T} (B^T Z B)^{-1} G^{-1} G B^T \\ &= B(B^T Z B)^{-1} B^T = Q. \end{aligned} \tag{17}$$

Finally, we mention the special case when $k = d$ and B is of full rank. From the argument in Eq. (17), we can take $B = I$ without loss of generality, which gives $Q = Z^{-1}$. This is the ordinary RX anomaly detector[16] in the stacked space.

4. ORDINARY LEAST SQUARES REGRESSION

For the chronochrome algorithm,[1, 2] the task is to find a matrix $L \in \mathbb{R}^{d_y \times d_x}$ for which $\hat{\mathbf{y}} = L\mathbf{x}$ approximates $\mathbf{y}$ in the least squares sense. That is,

$$\begin{aligned} L &= \operatorname{argmin}_L \left\langle ||\mathbf{y} - \hat{\mathbf{y}}||_F^2 \right\rangle \\ & \quad \text{such that } \hat{\mathbf{y}} = L\mathbf{x} \end{aligned} \tag{18}$$

where the $||\cdot||_F$ symbol indicates the Frobenius norm (square root of the sum of the squares of the elements of the matrix). Equivalently, we can say $\hat{D}_y = L D_x$ approximates D_y in the least squares (or Frobenius) sense; that is,

$$L = \operatorname{argmin}_L \, ||D_y - L D_x||_F. \tag{19}$$

We note that

$$\begin{aligned} ||D_y - L D_x||_F^2 &= \texttt{trace}\big([D_y - L D_x][D_y - L D_x]^T\big) && (20) \\ &= \texttt{trace}\big([D_y D_y^T - L D_x D_y^T - D_y D_x^T L^T + L D_x D_x^T L^T\big) && (21) \\ &= n \, \texttt{trace}\big(Y - L C^T - C L^T + L X L^T\big) && (22) \end{aligned}$$

is quadratic in L and in particular can be expressed in terms of the completed square:

$$Y - LC^T - CL^T + LXL^T = Y - CX^{-1}C^T + [LX^{1/2} - CX^{-1/2}][LX^{1/2} - CX^{-1/2}]^T. \tag{23}$$

Therefore

$$\begin{aligned}(1/n)||D_y - LD_x||_F^2 &= \texttt{trace}\big(Y - CX^{-1}C^T\big) + \texttt{trace}\Big([LX^{1/2} - CX^{-1/2}][LX^{1/2} - CX^{-1/2}]^T\Big) &(24)\\ &= \texttt{trace}\big(Y - CX^{-1}C^T\big) + ||LX^{1/2} - CX^{-1/2}||_F^2. &(25)\end{aligned}$$

It is clear that this achieves its minimum when $LX^{1/2} - CX^{-1/2} = 0$, or equivalently, when $L = CX^{-1}$.

Following the framework in Fig. 1, we consider the residuals from the regressed values. Thus,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - CX^{-1}\mathbf{x} \tag{26}$$

with larger residuals corresponding to more anomalous changes. To convert the vector-valued $\mathbf{e}$ to a scalar anomalousness, we use the Mahalanobis distance:

$$\mathcal{A}_{\rm CC}(\mathbf{x},\mathbf{y}) = \mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e} = (\mathbf{y} - CX^{-1}\mathbf{x})^T \left[Y - CX^{-1}C^T\right]^{-1} (\mathbf{y} - CX^{-1}\mathbf{x}). \tag{27}$$

This is a quadratic detector, $\mathcal{A}_{\rm CC}(\mathbf{x},\mathbf{y}) = \mathbf{z}^T Q_{\rm CC}\mathbf{z}$, where $\mathbf{z}$ is defined in Eq. (4), and $Q_{\rm CC}$ is given by

$$\begin{aligned}Q_{\rm CC} &= \begin{bmatrix} -X^{-1}C^T \\ I \end{bmatrix} \left[Y - CX^{-1}C\right]^{-1} \begin{bmatrix} -CX^{-1} & I \end{bmatrix} &(28)\\ &= \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}^{-1} - \begin{bmatrix} X^{-1} & 0 \\ 0 & 0 \end{bmatrix}. &(29)\end{aligned}$$

where the equivalence of Eq. (28) and Eq. (29) was shown in Ref. [13].

4.1. Asymmetry of the chronochrome

The chronochrome is inherently asymmetric with respect to x and y. Eq. (27) was derived by regressing $\mathbf{y}$ against $\mathbf{x}$, but a second chronochrome can be obtained by regressing $\mathbf{x}$ against $\mathbf{y}$. Here, we write $\hat{\mathbf{x}} = L'\mathbf{y}$ and choose L' to minimize $\langle ||\mathbf{x} - \hat{\mathbf{x}}||_F \rangle$, or equivalently

$$L' = \text{argmin}_L \; ||D_x - LD_y||_F, \tag{30}$$

for which the solution is $L' = C^T Y^{-1}$. The anomaly detector is given by

$$\mathcal{A}_{\rm CC'}(\mathbf{x},\mathbf{y}) = (\mathbf{x} - C^TY^{-1}\mathbf{y})^T \left[X - C^TY^{-1}C\right]^{-1} (\mathbf{x} - C^TY^{-1}\mathbf{y}). \tag{31}$$

and in general, the expressions for anomalousness in Eq. (27) and Eq. (31) are not equivalent. In fact, Eq. (31) corresponds to Eq. (12) with

$$Q_{\rm CC'} = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & Y^{-1} \end{bmatrix}. \tag{32}$$

4.2. Scale invariance of the chronochrome

The anomalies found by (both variants of) chronochrome are invariant to scaling of the data. Specifically, consider linearly scaled datasets $D'_x = G_xD_x$ and $D'_y = G_yD_y$, where G_x and G_y are nonsingular matrices. For these transformed data, following Eqs. (7-10), we have $X' = G_xXG_x^T$, $Y' = G_yYG_y^T$, and $C' = G_yCG_x^T$. Further, consider quadratic coefficient matrices of the form

$$Q = \begin{bmatrix} X & C^T \\ C & Y \end{bmatrix}^{-1} - \begin{bmatrix} \beta_xX^{-1} & 0 \\ 0 & \beta_yY^{-1} \end{bmatrix}, \tag{33}$$

where appropriate choices of the scalar parameters β_x and β_y correspond to Eq. (29) or Eq. (32). Then, in the transformed coordinates, we have

$$\begin{aligned} Q' &= \begin{bmatrix} X' & C'^T \\ C' & Y' \end{bmatrix}^{-1} - \begin{bmatrix} \beta_x X'^{-1} & 0 \\ 0 & \beta_y Y'^{-1} \end{bmatrix} \\ &= \begin{bmatrix} G_x X G_x^T & G_y C G_x{}^T \\ G_y C G_x & G_y Y G_y^T \end{bmatrix}^{-1} - \begin{bmatrix} \beta_x (G_x X G_x^T)^{-1} & 0 \\ 0 & \beta_y (G_y Y G_y^T)^{-1} \end{bmatrix} = G^{-T} Q G^{-1} \end{aligned} \tag{34}$$

where

$$G = \begin{bmatrix} G_x & 0 \\ 0 & G_y \end{bmatrix}. \tag{35}$$

Thus, the anomalousness in transformed coordinates is the same as the anomalousness in original coordinates:

$$\mathcal{A}'(\mathbf{x}', \mathbf{y}') = \mathbf{z}'^T Q' \mathbf{z}' = \mathbf{z}^T G^T \left[G^{-T} Q G^{-1} \right] G \mathbf{z} = \mathbf{z} Q \mathbf{z} = \mathcal{A}(\mathbf{x}, \mathbf{y}). \tag{36}$$

5. TOTAL LEAST SQUARES

The concept of total least squares was popularized by an important paper of Golub and Van Loan.[11] For total least squares, we allow errors in both $\mathbf{x}$ and $\mathbf{y}$. We write $\hat{\mathbf{y}} = L\hat{\mathbf{x}}$ and choose $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$ to minimize the total error:

$$\begin{aligned} L &= \operatorname{argmin}_L \left\langle ||\mathbf{y} - \hat{\mathbf{y}}||_F^2 + ||\mathbf{x} - \hat{\mathbf{x}}||_F^2 \right\rangle. \\ & \quad \text{such that } \hat{\mathbf{y}} = L\hat{\mathbf{x}}. \end{aligned} \tag{37}$$

Note that, in terms of the stacked variable $\mathbf{z}$, we have $||\mathbf{y} - \hat{\mathbf{y}}||_F^2 + ||\mathbf{x} - \hat{\mathbf{x}}||_F^2 = ||\mathbf{z} - \hat{\mathbf{z}}||_F^2$, and that $\hat{\mathbf{y}} = L\hat{\mathbf{x}}$ is equivalent to $K\hat{\mathbf{z}} = 0$ where $K = [-L \mid I]$; that is: the first d_x columns of K correspond to the columns of the matrix $-L$ and the last d_y columns are the identity matrix. This suggests a generalization of Eq. (37):

$$\begin{aligned} K &= \operatorname{argmin}_K \left\langle ||\mathbf{z} - \hat{\mathbf{z}}||_F^2 \right\rangle \\ & \quad \text{such that } K\hat{\mathbf{z}} = 0, \text{ and} \\ & \quad \operatorname{rank}(K) \geq k. \end{aligned} \tag{38}$$

The condition on the rank of K is important; among other things, it rules out the trivial minimum given by $K = 0$ and $\hat{\mathbf{z}} = \mathbf{z}$. Here, K plays the role of the matrix $[-L \mid I]$ in the formulation in Eq. (37), but as expressed in Eq. (38), it is not uniquely defined. For instance, if K solves Eq. (38), then for any nonsingular matrix G with as many columns as K has rows, GK also solves Eq. (38).

The case where $\operatorname{rank}(K) = d_y$ is instructive. If K is a rank d_y matrix, then the nullspace condition $K\hat{\mathbf{z}} = 0$ can equivalently be expressed with a matrix K' of size $d_y \times d$. (Since K is of rank d_y, it must have d_y linearly independent rows: construct K' by selecting these d_y rows from K.) Write this matrix $K' = [K_a \mid K_b]$ where K_a is a $d_y \times (d - d_y)$ matrix and K_b is $d_y \times d_y$. As long as K_b is invertible, then $K\hat{\mathbf{z}} = 0$ is equivalent to $K_b^{-1} K' \hat{\mathbf{z}} = 0$. But note that $K_b^{-1} K' = \left[K_b^{-1} K_a \mid I \right]$. If we now write $L = -K_b^{-1} K_a$, then we have $[-L \mid I]\, \hat{\mathbf{z}} = 0$, or equivalently $\hat{\mathbf{y}} = L\hat{\mathbf{x}}$. In other words, solving the more general problem in Eq. (38) with the constraint that $\operatorname{rank}(K) = d_y$ is equivalent to solving the problem in Eq. (37).

In a more modern interpretation of total least squares,[12] the problem is to find a lower-rank approximation of the data matrix. Applied to our $d = d_x + d_y$ dimensional stacked vector data D_z, we write

$$D_z^{(m)} = \operatorname{argmin}_D ||D_z - D||_F \text{ such that } \operatorname{rank}(D) \leq m \tag{39}$$

We can explicitly write the solution to this optimization in terms of a singular value decomposition. Let $D_z = U \Sigma V^T$ be the singular value decomposition of the data matrix. Here U and V are orthogonal matrices (satisfying $U^T U = I$ and $V^T V = I$), and Σ is a diagonal matrix of non-negative singular values. This enables us to write $D_z^{(m)} = U \Sigma_m V^T$ where Σ_m is the best rank-m approximation to Σ; in particular, Σ_m is Σ with the $d - m$ smallest elements set to zero (thus, the largest m elements are retained).

Note that we do not have to perform a singular value decomposition on the entire dataset D_z, but can instead obtain U and Σ directly from the covariance matrix Z (or, equivalently, from X, Y and C). In particular, $Z = (1/n)D_zD_z^T = (1/n)U\Sigma^2U^T$. So U is the matrix of eigenvectors of Z, and $(1/n)\Sigma^2$ is a diagonal matrix of eigenvalues.

We remark that $\Delta_m = \Sigma_m\Sigma^{-1} = \Sigma^{-1}\Sigma_m$ is a diagonal matrix with only ones and zeros on the diagonal (in fact, it has m ones, and $d-m$ zeros), and that it is therefore idempotent: that is, $\Delta_m^2 = \Delta_m$. In terms of U and Δ_m, we can write the projection operator $P_m = U\Delta_mU^T$, and observe that $D_z^{(m)} = P_mD_z$ or, on a pixel-by-pixel basis, $\hat{\mathbf{z}} = P_m\mathbf{z}$. Note that P_m is also idempotent and that the rank of P_m is m. Note further that $(I - P_m)$ is a projection; it is orthogonal to P_m since $(I-P_m)P_m = P_m - P_m^2 = P_m - P_m = 0$, and its rank is $d-m$.

This enables us to solve Eq. (38). Let $K = (I - P_m)$ and observe that

$$K\hat{\mathbf{z}} = (I - P_m)\hat{\mathbf{z}} = (I - P_m)P_m\hat{\mathbf{z}} = 0. \tag{40}$$

The rank of K is $d-m$, so we can take $m = d-k$ to ensure that the rank of K is at least k and that both conditions in Eq. (38) are therefore satisfied. Finally, as explained above, the case $k = d_y$ can be used to solve Eq. (37).

5.1. Total least squares for anomalous change detection

Consider the residuals of this approximation, $\mathbf{e} = \mathbf{z} - \hat{\mathbf{z}}$. Since $\hat{\mathbf{z}} = P_{d-k}\mathbf{z}$, we can write $\mathbf{e} = (I - P_{d-k})\mathbf{z}$. Observe that $I - P_{d-k} = I - U\Delta_{d-k}U^T = U(I - \Delta_{d-k})U^T$, and note that $I - \Delta_{d-k}$ is a diagonal matrix with ones and zeros on the diagonal; in fact it has ones where Δ_{d-k} has zeros and vice versa. Since Δ_{d-k} has ones corresponding to the $d-k$ largest eigenvalues of Z, it follows that $I - \Delta_{d-k}$ will have ones corresponding to the k smallest eigenvalues of Z. Let B_k be the k nonzero columns of $U(I - \Delta_{d-k})$; these correspond to the k eigenvectors of Z with the smallest eigenvalues. Further, we can write $(I - P_{d-k}) = U(I - \Delta_{d-k})U^T = U(I - \Delta_{d-k})^2U^T = B_kB_k^T$, and so

$$\left\langle \mathbf{e}\mathbf{e}^T \right\rangle = \left\langle (I - P_{d-k})\mathbf{z}\mathbf{z}^T(I - P_{d-k}^T) \right\rangle = B_kB_k^TZB_kB_k^T \tag{41}$$

the *pseudoinverse* of which is given by $B_k(B_k^TZB_k)^{-1}B_k^T$. It follows that the anomalies are given by

$$\mathcal{A}_{\mathrm{TLSQ}}(\mathbf{z}) = \mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e} = \mathbf{z}^TB_kB_k^TB_k(B_k^TZB_k)^{-1}B_k^TB_kB_k^T\mathbf{z} = \mathbf{z}^TB_k(B_k^TZB_k)^{-1}B_k^T\mathbf{z} = \mathbf{z}^TQ_{\mathrm{TLSQ}}\mathbf{z} \tag{42}$$

with

$$Q_{\mathrm{TLSQ}} = B_k(B_k^TZB_k)^{-1}B_k^T, \tag{43}$$

where, again, the columns of B_k are the k eigenvectors of Z with the smallest eigenvalues. We remark that this Q is the best rank-k approximation to Z^{-1}; that is,

$$Q_{\mathrm{TLSQ}} = \mathrm{argmin}_Q ||Z^{-1} - Q||_F \text{ such that } \mathrm{rank}(Q) \le k \tag{44}$$

We also remark that the TLSQ detector is the same as the subspace RX (SSRX) anomaly detector,[17] but it is used here to detect anomalous *changes* by looking for anomalies in the stacked space.

5.2. Lack of scale invariance for total least squares

The argument in Section 4.2 does not directly apply for TLSQ because Q cannot be expressed in a form given by Eq. (33). Nonetheless following the argument in Section 4.2, we write $Z' = GZG^T$ and let B'_k be the k eigenvectors of Z'. In these coordinates, the coefficient matrix would look like Eq. (43): $Q' = B'_k({B'}_k^TZ'B'_k)^{-1}{B'}_k^T$. But expressed in the original coordinates, we get

$$Q_{G-TLSQ} = G^TQ'G = G^TB'_k({B'}_k^TGZG^TB'_k)^{-1}{B'}_k^TG. \tag{45}$$

Comparing this with Eq. (43), we have equality if $G^TB'_k = B_k$, but that's not generally the case. $G^TB'_k$ is the transformed eigenvectors of the transformed covariance Z', while B_k is the original eigenvectors of the original covariance Z.

This lack of scale invariance in TLSQ (which is also lacking in SSRX) leads us to specify a standardized scaling of the data.

5.3. Whitened total least squares

A natural way to deal with the lack of scale invariance is to individually whiten the $\mathbf{x}$ and $\mathbf{y}$ images before TLSQ is applied. In these coordinates, we write

$$\tilde{\mathbf{x}} = X^{-1/2}\mathbf{x} \tag{46}$$

$$\tilde{\mathbf{y}} = Y^{-1/2}\mathbf{y} \tag{47}$$

$$\tilde{\mathbf{z}} = \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} X^{-1/2} & 0 \\ 0 & Y^{-1/2} \end{bmatrix} \mathbf{z}. \tag{48}$$

Note that the x-image and y-image are *not collectively* whitened; so $\tilde{\mathbf{z}}$ is *not* given by $Z^{-1/2}\mathbf{z}$. The covariances in the whitened coordinates are simplified:

$$\tilde{X} = \left\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \right\rangle = I \tag{49}$$

$$\tilde{Y} = \left\langle \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T \right\rangle = I \tag{50}$$

$$\tilde{C} = \left\langle \tilde{\mathbf{y}}\tilde{\mathbf{x}}^T \right\rangle = Y^{-1/2} C X^{-1/2} \tag{51}$$

$$\tilde{Z} = \left\langle \tilde{\mathbf{z}}\tilde{\mathbf{z}}^T \right\rangle = \begin{bmatrix} I & \tilde{C}^T \\ \tilde{C} & I \end{bmatrix} = I + \begin{bmatrix} 0 & \tilde{C}^T \\ \tilde{C} & 0 \end{bmatrix}. \tag{52}$$

Using $\tilde{z}$ defined in Eq. (48) and $\tilde{Z}$ defined in Eq. (52), we follow the derivation for TLSQ above, but in whitened coordinates, to produce an anomaly detector

$$\mathcal{A}_{\mathrm{WTLSQ}}(\tilde{\mathbf{z}}) = \tilde{\mathbf{z}}^T \tilde{Q}_{\mathrm{WTLSQ}} \tilde{\mathbf{z}} \tag{53}$$

with

$$\tilde{Q}_{\mathrm{WTLSQ}} = \tilde{B}_k (\tilde{B}_k^T \tilde{Z} \tilde{B}_k)^{-1} \tilde{B}_k^T \tag{54}$$

where $\tilde{B}_k$ corresponds to the k smallest eigenvectors of $\tilde{Z}$.

We can express this detector in the original (unwhitened) coordinates, using $Q_{\mathrm{WTLSQ}} = G_w^T \tilde{Q}_{\mathrm{WTLSQ}} G_w$, where

$$G_w = \begin{bmatrix} X^{-1/2} & 0 \\ 0 & Y^{-1/2} \end{bmatrix} \tag{55}$$

is the whitening transform. Recognizing that $\tilde{Z} = G_w Z G_w^T$, and writing $\breve{B}_k = G_w^T \tilde{B}_k$, then we obtain

$$Q_{\mathrm{WTLSQ}} = G_w^T \tilde{B}_k (\tilde{B}_k^T G_w Z G_w^T \tilde{B}_k)^{-1} \tilde{B}_k^T G_w = \breve{B}_k (\breve{B}_k^T Z \breve{B}_k)^{-1} \breve{B}_k^T \tag{56}$$

which evokes Eq. (43), but is not the same because $\breve{B}_k$ is not the same as the B_k that appears in Eq. (43). The whitened eigenvectors of the whitened covariance $\tilde{Z}$ is not the same as the original eigenvectors of the original covariance Z.

In the following sections, we will show that (for some values of k), the whitened total least squares anomalous change detector is equivalent to the use of canonical correlation analysis in the multivariate alteration detection (MAD) algorithm[6] and to the optimized covariance equalization algorithm.[4, 5]

5.4. Comparison to Canonical Correlation Analysis

Nielsen *et al.*[6] introduced an ACD algorithm called multivariate alteration detection (MAD), based on canonical correlation analysis (CCA). The aim of CCA is to find linear combinations of data that are maximally correlated to each other.[7] This is most easily understood for the first canonical correlation component. We want to choose vectors $\mathbf{b}_x \in \mathbb{R}^{d_x}$ and $\mathbf{b}_y \in \mathbb{R}^{d_y}$ so that the scalar values $x = \mathbf{b}_x^T\mathbf{x}$ and $y = \mathbf{b}_y^T\mathbf{y}$ are highly correlated. Specifically we want to maximize:

$$\rho = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}} = \frac{\mathbf{b}_x^T C^T \mathbf{b}_y}{\sqrt{(\mathbf{b}_x^T X \mathbf{b}_x)(\mathbf{b}_y^T Y \mathbf{b}_y)}}. \tag{57}$$

Because the magnitudes of $\mathbf{b}_x$ and $\mathbf{b}_y$ don't affect this correlation, we can impose constraints

$$\begin{aligned} 1 &= \langle x^2 \rangle = \langle (\mathbf{b}_x^T \mathbf{x})^2 \rangle = \mathbf{b}_x^T \langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{b}_x = \mathbf{b}_x^T X \mathbf{b}_x & (58) \\ 1 &= \langle y^2 \rangle = \langle (\mathbf{b}_y^T \mathbf{y})^2 \rangle = \mathbf{b}_y^T \langle \mathbf{y}\mathbf{y}^T \rangle \mathbf{b}_y = \mathbf{b}_y^T Y \mathbf{b}_y & (59) \end{aligned}$$

on the magnitudes of $\mathbf{b}_x^T\mathbf{x}$ and $\mathbf{b}_y^T\mathbf{y}$. With these in place, we want to maximize the correlation, which is given by the product $\langle (\mathbf{b}_x^T\mathbf{x})(\mathbf{b}_y^T\mathbf{y}) \rangle = \mathbf{b}_x^T \langle \mathbf{x}\mathbf{y}^T \rangle \mathbf{b}_y = \mathbf{b}_x^T C^T \mathbf{b}_y$. Subject to these constraints, we can write

$$\rho = \mathbf{b}_x^T C^T \mathbf{b}_y - \mu(\mathbf{b}_x^T X \mathbf{b}_x - 1) - \nu(\mathbf{b}_y^T Y \mathbf{b}_y - 1) \tag{60}$$

where μ and ν are Lagrange multipliers. Differentiating with respect to $\mathbf{b}_x$ and $\mathbf{b}_y$, and setting to zero, we obtain:

$$\begin{aligned} 0 &= \frac{\partial \rho}{\partial \mathbf{b}_x} = C^T \mathbf{b}_y - 2\mu X \mathbf{b}_x, & (61) \\ 0 &= \frac{\partial \rho}{\partial \mathbf{b}_y} = C \mathbf{b}_x - 2\nu Y \mathbf{b}_y. & (62) \end{aligned}$$

Multiplying Eq. (61) by $\mathbf{b}_x^T$ and Eq. (62) by $\mathbf{b}_y^T$:

$$\begin{aligned} 0 &= \mathbf{b}_x^T C^T \mathbf{b}_y - 2\mu \mathbf{b}_x^T X \mathbf{b}_x = \mathbf{b}_x^T C^T \mathbf{b}_y - 2\mu, & (63) \\ 0 &= \mathbf{b}_y^T C \mathbf{b}_x - 2\nu \mathbf{b}_y^T Y \mathbf{b}_y = \mathbf{b}_y^T C \mathbf{b}_x - 2\nu. & (64) \end{aligned}$$

which gives $\mu = \nu = (1/2)\mathbf{b}_x^T C^T \mathbf{b}_y$. Let $\lambda = \mathbf{b}_x^T C^T \mathbf{b}_y = 2\mu = 2\nu$ and rewrite Eq. (61) and Eq. (62) as

$$\begin{bmatrix} 0 & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{b}_x \\ \mathbf{b}_y \end{bmatrix} - \lambda \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \begin{bmatrix} \mathbf{b}_x \\ \mathbf{b}_y \end{bmatrix} = 0. \tag{65}$$

If we write $\tilde{\mathbf{b}}_x = X^{1/2}\mathbf{b}_x$ and $\tilde{\mathbf{b}}_y = Y^{1/2}\mathbf{b}_y$, then this simplifies to

$$\begin{bmatrix} 0 & \tilde{C}^T \\ \tilde{C} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix} = \lambda \begin{bmatrix} \tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix}. \tag{66}$$

where $\tilde{C} = Y^{-1/2} C X^{-1/2}$. This expression can be rewritten

$$\begin{bmatrix} I & \tilde{C}^T \\ \tilde{C} & I \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix} = (1+\lambda) \begin{bmatrix} \tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix}. \tag{67}$$

and we see that the whitened form of the first canonical component $(\tilde{\mathbf{b}}_x, \tilde{\mathbf{b}}_y)$ is an eigenvector of the whitened covariance matrix $\tilde{Z}$ defined in Eq. (52). The first canonical component is given by the eigenvector with largest eigenvalue; subsequent canonical components are given by eigenvectors with decreasing eigenvalues.

Because of the symmetry in $\tilde{Z}$, there is a natural pairing of eigenvectors. With $\tilde{\mathbf{b}}_x$ and $\tilde{\mathbf{b}}_y$ chosen to satisfy Eq. (67), we can see

$$\begin{bmatrix} I & \tilde{C}^T \\ \tilde{C} & I \end{bmatrix} \begin{bmatrix} -\tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix} = (1-\lambda) \begin{bmatrix} -\tilde{\mathbf{b}}_x \\ \tilde{\mathbf{b}}_y \end{bmatrix}. \tag{68}$$

Write B_{kx} as the matrix whose k columns are the first k canonical components, and similarly for B_{ky}. Then the matrix

$$B_k = \begin{bmatrix} -B_{kx} \\ B_{ky} \end{bmatrix}, \tag{69}$$

will correspond to the k eigenvectors of $\tilde{Z}$ with the smallest eigenvalues.

Then using $\mathbf{x}' = B_{kx}^T \tilde{\mathbf{x}}$ and $\mathbf{y}' = B_{ky}^T \tilde{\mathbf{y}}$ as new coordinates, we can look for anomalous changes by just subtracting them

$$\mathbf{e} = \mathbf{y}' - \mathbf{x}' = B_{ky}^T \tilde{\mathbf{y}} - B_{kx}^T \tilde{\mathbf{x}} = B_k \mathbf{z} \tag{70}$$

and expressing anomalousness as the Mahalanobis distance $\mathcal{A}(\mathbf{e}) = \mathbf{e}^T \left\langle \mathbf{e}\mathbf{e}^T \right\rangle^{-1} \mathbf{e}$, which gives $\mathcal{A}_{\mathrm{CCA}}(\tilde{\mathbf{z}}) = \tilde{\mathbf{z}}^T Q_{\mathrm{CCA}} \tilde{\mathbf{z}}$ where

$$\tilde{Q}_{\mathrm{CCA}} = B_k (B_k^T \tilde{Z} B_k)^{-1} B_k^T. \tag{71}$$

with B_k corresponding to the k smallest eigenvectors of $\tilde{Z}$. Comparing Q_{CCA} to Q_{WTLSQ} in Eq. (54) shows that the CCA-based anomaly detector is identical to the whitened TLSQ detector.

In recommending canonical correlation analysis for the multivariate alteration detection (MAD) algorithm,[6] the role of k was not highlighted. Since typically, $d_x = d_y$, it is implicitly recommended that $k = d_x = d_y$. But the formalism permits values of k ranging from $k = 1$ to $k = d = d_x + d_y$, with the latter case corresponding to the RX algorithm applied to the stacked vector $\mathbf{z}$. When $k < \min(d_x, d_y)$, then CCA can be treated as a dimension reduction algorithm, reducing both the x-image and y-image to dimension k. These dimension-reduced images can then be used as input images for other ACD algorithms, quadratic[13] or otherwise.[18]

5.5. Comparison to Optimized Covariance Equalization

The idea of covariance equalization is, as the name suggests, to transform the x-image and y-image so that they share the same covariance. Then the images are subtracted and anomalousness is computed in terms of Mahalanobis distance of the vector-valued difference.

If $\mathbf{x}' = B_x^T \mathbf{x}$, then $\left\langle \mathbf{x}'\mathbf{x}'^T \right\rangle = B_x^T \left\langle \mathbf{x}\mathbf{x}^T \right\rangle B_x = B_x^T X B_x$ is the covariance of the transformed x-image. Similarly, $B_y^T Y B_y$ is the covariance of the transformed y-image. In the covariance equalization scheme, then, we choose B_x and B_y so that

$$B_x^T X B_x = B_y^T Y B_y. \tag{72}$$

For "standard" covariance equalization, we take $B_x = X^{-1/2}$ and $B_y = Y^{-1/2}$. In this case, $\mathbf{x}'$ and $\mathbf{y}'$ are both white (their covariance matrices are the identity matrix). An alternative is to take $B_x = X^{-1/2} Y^{1/2}$ and $B_y = I$, so $\mathbf{x}'$ and $\mathbf{y}'$ will both have the same covariance as the original y-image. Since this corresponds to a linear invertible transform to the matrix $B = \begin{bmatrix} -B_x & \\ & B_y \end{bmatrix}$, we have from Eq. (17) that the result will be the same. We further remark that standard covariance equalization requires that $d_x = d_y$.

A more general solution to Eq. (72) is given by $B_x = X^{-1/2} R$ and $B_y = Y^{-1/2} S$ where R and S satisfy $R^T R = I$ and $S^T S = I$. It is not necessary that R or S be square, so the case $d_x \neq d_y$ can be accommodated. Specifically, we can say that R is a $d_x \times k$ matrix and S is $d_y \times k$.

The "optimal" covariance equalization chooses R and S to minimize average squared difference of the transformed variables, $\langle ||B_y^T \mathbf{y} - B_x^T \mathbf{x}||_F^2 \rangle$. In particular, we note

$$\begin{aligned} \langle ||B_y^T \mathbf{y} - B_x^T \mathbf{x}||_F^2 \rangle &= \mathtt{trace}\big(B_y^T Y B_y - B_x^T C^T B_y - B_y^T C B_x + B_x^T X B_x\big) && (73) \\ &= \mathtt{trace}\Big(I - R^T \tilde{C}^T S - S^T \tilde{C} R + I\Big) && (74) \end{aligned}$$

so that the choice of R and S should maximize $\mathtt{trace}\Big(S^T \tilde{C} R\Big)$.

Let $\tilde{C} = U J V^T$ be its singular value decomposition. Recall that $\tilde{C}$ is a matrix of size $d_y \times d_x$, and so it has at most $m = \min(d_x, d_y)$ nonzero singular values. Thus, we can write the decomposition with U and V both having m columns, and J being a diagonal $m \times m$ matrix.

In the case for which covariance equalization was originally proposed, $k = m = d_x = d_y$, and $\mathtt{trace}\Big(S^T \tilde{C} R\Big) = \mathtt{trace}\big(S^T U J V^T R\big) = \mathtt{trace}\big(V^T R S^T U J\big) \leq \mathtt{trace}(J)$ with equality obtained when $V^T R S^T U = I$, or $S R^T = U V^T = (\tilde{C}\tilde{C}^T)^{-1/2} \tilde{C}$. The particular choice $R^T = U V^T$ and $S = I$ was proposed by Schaum and Stocker.[5]

If we more generally assume $k \leq m$, then $S^T \tilde{C} R$ will be a $k \times k$ matrix, and $\mathtt{trace}\big(S^T U J V^T R\big) \leq \sum_{n=1}^k j_n$ where j_n is the nth singular value. We can achieve this bound with $S R^T = U_k V_k^T$ where U_k corresponds to the first k columns of U and V_k is the first k columns of V. The choice $S = U_k$ and $R = V_k$ is a particularly convenient formulation and leads to "diagonalized" covariance equalization.[13]

This leads to the transform

$$B_k = \begin{bmatrix} -R \\ S \end{bmatrix} = \begin{bmatrix} -V_k \\ U_k \end{bmatrix} \tag{75}$$

which is a $d \times k$ matrix, the nth column of which is given by

$$\mathbf{b}_n = \begin{bmatrix} -\mathbf{v}_n \\ \mathbf{u}_n \end{bmatrix} \tag{76}$$

where $\mathbf{v}_n$ (resp. $\mathbf{u}_n$) is the nth column of V (resp. U).

In general we note that

$$\begin{aligned} \tilde{Z} \begin{bmatrix} \pm\mathbf{v}_n \\ \mathbf{u}_n \end{bmatrix} &= \begin{bmatrix} I & \tilde{C}^T \\ \tilde{C} & I \end{bmatrix} \begin{bmatrix} \pm\mathbf{v}_n \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} I & VJU^T \\ UJV^T & I \end{bmatrix} \begin{bmatrix} \pm\mathbf{v}_n \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \pm\mathbf{v}_n + VJU^T\mathbf{u}_n \\ \mathbf{u}_n \pm UJV^T\mathbf{v}_n \end{bmatrix} \\ &= \begin{bmatrix} \pm\mathbf{v}_n + j_n\mathbf{v}_n \\ \mathbf{u}_n \pm j_n\mathbf{u}_n \end{bmatrix} = (1 \pm j_n) \begin{bmatrix} \pm\mathbf{v}_n \\ \mathbf{u}_n \end{bmatrix} \end{aligned} \tag{77}$$

In terms of the $m = \min(d_x, d_y)$ singular values j_n of $\tilde{C}$, we can say that the m largest eigenvalues of $\tilde{Z}$ are of the form $1 + j_n$, that the m smallest are of the form $1 - j_n$, and the rest are equal to 1. And the smallest ones are associated with eigenvectors of the form in Eq. (76). It follows that the the coefficient matrix for optimized covariance equalization is given by

$$\tilde{Q}_{\text{CE-R}} = B_k(B_k^T \tilde{Z} B_k)^{-1} B_k^T \tag{78}$$

where the columns of B_k are the k smallest eigenvectors of $\tilde{Z}$, and therefore the detector defined by Eq. (78) is identical to the WTLSQ detector in Eq. (54) and the CCA detector in Eq. (71). We remark that this formulation requires $k \leq m = \min(d_x, d_y)$, but that the more general whitened TLSQ formulation permits the range $1 \leq k \leq d = d_x + d_y$.

6. DISTRIBUTION-BASED CHANGE DETECTION

Another way to think about anomalous change detection is as a binary detection problem.[8] Write $P(\mathbf{x}, \mathbf{y})$ is the joint distribution of pixel values $(\mathbf{x}, \mathbf{y})$ in the two images. Write $P_a(\mathbf{x}, \mathbf{y})$ as a "model" for anomalous pixels. It may seem counter-intuitive to write such an explicit model for something that is known to defy definition. Anomalies, after all, are the ultimate *je ne sais quoi*: we say that they are *ir*regular or *un*common or *a*typical, but we become more equivocal when we try to say positively what they are. As long as $P_a(\mathbf{x}, \mathbf{y})$ is a broad and diffuse distribution, however, then the anomalies that it "defines" will maintain their open-ended character. The advantage of specifying $P_a(\mathbf{x}, \mathbf{y})$ explicitly is that it leads to a likelihood ratio $P(\mathbf{x}, \mathbf{y})/P_a(\mathbf{x}, \mathbf{y})$ which is the unambiguously optimal solution to the anomaly detection problem. Turning this around: any detector which can be expressed as a ratio $P(\mathbf{x}, \mathbf{y})/P_a(\mathbf{x}, \mathbf{y})$ can be interpreted as the optimal detector of anomalies whose nature is specified by $P_a(\mathbf{x}, \mathbf{y})$.

In particular, if we can express a quadratic anomalous change detection algorithm as $Q = Z^{-1} - W^{-1}$ where Z is defined in Eq. (5) and W^{-1} is positive semi-definite, then we can treat W as a covariance matrix that describes the "background distribution" $P_a(\mathbf{x}, \mathbf{y})$ for anomalies. For RX, $W^{-1} = 0$ which corresponds to a uniform background (*i.e.,* a Gaussian with infinite variance in all directions). For chronochrome, this difference is expressed in Eq. (29) and Eq. (32). For TLSQ (and, equivalently (in whitened coordinates), MAD and CE), $W^{-1} = U_{d-k}(U_{d-k}^T Z U_{d-k}^T)^{-1} U_{d-k}^T$, where U_{d-k} corresponds to the $d - k$ largest eigenvectors of Z.

7. CONCLUSIONS

Inspired by the chronochrome, an anomalous change detector based on ordinary least squares, we employ total least squares (TLSQ) to derive a new ACD algorithm. We find that the TLSQ-based anomalous change detector is equivalent to SSRX applied to the stacked image space. Unlike the chronochrome, however, TLSQ is not invariant to coordinate changes in x and y. To deal with this lack of invariance, we introduced whitened TLSQ, which is TLSQ applied to data in which the two images have been individually whitened.

In the whitened TLSQ algorithm, there is a user-chosen parameter k which corresponds to the rank of the quadratic coefficient matrix. For appropriately chosen values of k, whitened TLSQ is equivalent to two known ACD algorithms: multivariate alteration detection and optimized covariance equalization. The whitened TLSQ algorithm however exhibits more flexibility because the parameter k can range from 1 to d, the sum of the number of spectral channels in the x and y images. The case $k = d$ turns out to be equivalent to the RX detector applied to the stacked image space.

All of these algorithms are derived from a general subtraction-based change detection framework. But we provide an alternative interpretation in terms of a distribution-based framework.

ACKNOWLEDGMENTS

The support of the U.S. Department of Energy, through the Los Alamos Laboratory Directed Research and Development (LDRD) Program, is gratefully acknowledged.

REFERENCES

1. A. Schaum and A. Stocker, "Spectrally selective target detection," *Proc. ISSSR (International Symposium on Spectral Sensing Research)* , 1997.
2. A. Schaum and A. Stocker, "Long-interval chronochrome target detection," *Proc. ISSSR (International Symposium on Spectral Sensing Research)* , 1998.
3. C. Clifton, "Change detection in overhead imagery using neural networks," *Applied Intelligence* **18**, pp. 215–234, 2003.
4. A. Schaum and A. Stocker, "Linear chromodynamics models for hyperspectral target detection," *Proc. IEEE Aerospace Conference* , pp. 1879–1885, 2003.
5. A. Schaum and A. Stocker, "Estimating hyperspectral target signature evolution with a background chromodynamics model," *Proc. International Symposium on Spectral Sensing Research (ISSSR)* , 2003.
6. 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**, pp. 1–19, 1998.
7. H. Hotelling, "Relations between two sets of variables," *Biometrika* **28**, pp. 321–327, 1936.
8. 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.
9. J. Theiler and S. Perkins, "Resampling approach for anomalous change detection," *Proc. SPIE* **6565**, p. 65651U, 2007.
10. M. T. Eismann, J. Meola, A. D. Stocker, S. G. Beaven, and A. P. Schaum, "Airborne hyperspectral detection of small changes," *Applied Optics* **47**, pp. F27–F45, 2008.
11. G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numerical Analysis* **17**, pp. 883–893, 1980.
12. I. Markovsky and S. Van Huffel, "Overview of total least squares methods," *Signal Processing* **87**, pp. 2283–2302, 2007.
13. J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors," *Applied Optics* **47**, pp. F12–F26, 2008.
14. R. Penrose, "A generalized inverse for matrices," *Proc. Cambridge Philosophical Society* **51**, pp. 406–413, 1955.
15. G. Golub and W. Kanan, "Calculating the singular values and pseudo-inverse of a matrix," *J. SIAM Numerical Analysis, Series B* **2**, pp. 205–224, 1965.
16. 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.
17. A. Schaum, "Spectral subspace matched filtering," *Proc. SPIE* **4381**, pp. 1–17, 2001.
18. J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy, "Elliptically-contoured distributions for anomalous change detection in hyperspectral imagery," To appear in: *IEEE Geoscience and Remote Sensing Letters* , 2010. doi: 10.1109/LGRS.2009.2032565.

APPENDIX A. NOTATION AND LIST OF VARIABLES

In general, lower case italic font variables (*e.g.,* 'x') refer to scalars, lower case boldface (*e.g.,* '$\mathbf{x}$') to vectors, and upper case italic (*e.g.,* 'X') to matrices. The transpose of a matrix is indicated with a superscript T (*e.g.,* 'X^T'), and the matrix inverse (or pseudoinverse when the inverse doesn't exist) is indicated with a superscript -1 (*e.g.,* 'X^{-1}'). Since transpose of the inverse equals inverse of the transpose, we can use $-T$ to represent that case (*e.g.,* 'X^{-T}'). We have tried to explain each new variable in the text as it is introduced, but Table 1 summarizes the most commonly used variables. In some cases, *ad hoc* variables are introduced which may have different interpretations in different contexts.

Table 1. List of some commonly used variables

x (resp. y)	scalar value of a pixel in the x-image (resp. y-image)
$\mathbf{x}$ (resp. $\mathbf{y}$)	vector value of a multispectral pixel in the x-image (resp. y-image)
$\mathbf{z}$	vector value of a pixel in the "stacked" image, defined in Eq. (4)
d_x (resp. d_y)	dimension of x-image (resp. y-image)
$d = d_x + d_y$	dimension of stacked image
D_x (resp. D_y, D_z)	matrix of all the pixels in the x-image (resp. y-image, stacked image)
B_x (resp. B_y)	transformation applied to $\mathbf{x}$ (resp. $\mathbf{y}$)
$\mathbf{x}' = B_x^T\mathbf{x}$ (resp. $\mathbf{y}' = B_y^T\mathbf{x}$)	transformed vector value of a pixel in the x-image (resp. y-image)
B	transformation applied to stacked image $\mathbf{z}$
$\mathbf{e} = \mathbf{y}' - \mathbf{x}' = B\mathbf{z}$	vector-valued difference between transformed x-image and y-image
$\mathcal{A}$	scalar-valued anomalousness, usually a function of pixel value
X (resp. Y, Z)	covariance matrix of x-image (resp. y-image, stacked image)
C	cross-covariance of x-image and y-image, of size $d_y \times d_x$)
$\tilde{\mathbf{x}}$ (resp. $\tilde{\mathbf{y}}$, $\tilde{\mathbf{z}}$)	$\mathbf{x}$ (resp. $\mathbf{y}$, $\mathbf{z}$) in whitened coordinates
$\tilde{X}$ (resp. $\tilde{Y}$, $\tilde{C}$, $\tilde{Z}$)	matrix X (resp. Y, C, Z) in whitened coordinates
$\hat{\mathbf{x}}$ (resp. $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$)	approximation to $\mathbf{x}$ (resp. $\mathbf{y}$, $\mathbf{z}$)
G (resp. G_x, G_y)	invertible square matrix
G_w	invertible square matrix used as whitening transform, see Eq. (55).
R (resp. S)	possibly non-square matrix satisfying $R^TR = I$ (resp. $S^TS = I$)
L (resp. L')	linear map, of size $d_y \times d_x$ (resp. $d_x \times d_y$), used in chronochrome
K	rank-k matrix that defines TLSQ, defined in Eq. (38)
$D_z^{(m)}$	best rank-m approximation to data D_z
Q	matrix of coefficients for quadratic anomalousness: $\mathcal{A}(\mathbf{z}) = \mathbf{z}^TQ\mathbf{z}$
$\tilde{Q}$ (resp. Q')	coefficient matrix Q in whitened (resp. transformed) coordinates
U, V	matrix of eigenvectors, often obtained from singular value decomposition
U_k (resp. V_k)	submatrix of eigenvector matrix U (resp. V) with k columns corresponding to the k largest eigenvalues (or singular values, depending on context)
Σ	diagonal matrix of singular values of data D_z
Σ_m	best rank-m approximation to Σ; agrees with Σ for its m largest values
$\Delta_m = \Sigma^{-1}\Sigma_m$	diagonal matrix with m ones and the rest zeros
P_m	projection matrix that maps D_z into $D_z^{(m)}$ (equivalently, $\mathbf{z}$ into $\hat{\mathbf{z}}$)
$\breve{B}_k$	matrix used for expressing WTLSQ in original (unwhitened) coordinates
B_{kx} (resp. B_{ky}, B_k)	first k columns of B_x (resp. B_y, B)
J	diagonal matrix of singular values of $\tilde{C}$
j_n	nth singular value of $\tilde{C}$ (i.e., nth element of J)
$\mathbf{b}_n$ (resp. $\mathbf{u}_n$, $\mathbf{v}_n$)	vector obtained from the nth column of B (resp. U, V)
$P(\mathbf{x}, \mathbf{y})$	joint probability distribution for the pixel values $\mathbf{x}$ and $\mathbf{y}$
$P_a(\mathbf{x}, \mathbf{y})$	distribution for anomalous values $\mathbf{x}$ and $\mathbf{y}$
$W^{-1} = Q - Z^{-1}$	inverse covariance matrix of "background distribution"