The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi- and Hyperspectral Data

Allan Aasbjerg Nielsen

Abstract—This paper describes new extensions to the previously published multivariate alteration detection (MAD) method for change detection in bi-temporal, multi- and hypervariate data such as remote sensing imagery. Much like boosting methods often applied in data mining work, the iteratively reweighted (IR) MAD method in a series of iterations places increasing focus on "difficult" observations, here observations whose change status over time is uncertain. The MAD method is based on the established technique of canonical correlation analysis: for the multivariate data acquired at two points in time and covering the same geographical region, we calculate the canonical variates and subtract them from each other. These orthogonal differences contain maximum information on joint change in all variables (spectral bands). The change detected in this fashion is invariant to separate linear (affine) transformations in the originally measured variables at the two points in time, such as 1) changes in gain and offset in the measuring device used to acquire the data, 2) data normalization or calibration schemes that are linear (affine) in the gray values of the original variables, or 3) orthogonal or other affine transformations, such as principal component (PC) or maximum autocorrelation factor (MAF) transformations. The IR-MAD method first calculates ordinary canonical and original MAD variates. In the following iterations we apply different weights to the observations, large weights being assigned to observations that show little change, i.e., for which the sum of squared, standardized MAD variates is small, and small weights being assigned to observations for which the sum is large. Like the original MAD method, the iterative extension is invariant to linear (affine) transformations of the original variables. To stabilize solutions to the (IR-)MAD problem, some form of regularization may be needed. This is especially useful for work on hyperspectral data. This paper describes ordinary two-set canonical correlation analysis, the MAD transformation, the iterative extension, and three regularization schemes. A simple case with real Landsat Thematic Mapper (TM) data at one point in time and (partly) constructed data at the other point in time that demonstrates the superiority of the iterative scheme over the original MAD method is shown. Also, examples with SPOT High Resolution Visible data from an agricultural region in Kenya, and hyperspectral airborne HyMap data from a small rural area in southeastern Germany are given. The latter case demonstrates the need for regularization.

Index Terms—Canonical correlation analysis (CCA), iteratively reweighted multivariate alteration detection (IR-MAD), MAD transformation, regularization or penalization, remote sensing.

I. INTRODUCTION

THIS paper deals with detection of nontrivial change in multi- and hypervariate, bi-temporal data. The term "nontrivial" here means nonaffine change between two points in time. This means that changes—due to, for instance, 1) an additive shift in mean level (offset) or a multiplicative shift in calibration of a measuring device (gain), 2) data normalization or calibration schemes that are linear (affine) in the gray values of the original variables, or 3) orthogonal or other affine transformations such as principal component (PC) or maximum autocorrelation factor (MAF) transformations—are not detected. This invariance is an enormous advantage over most other multivariate change detection schemes published, see [1] for an early survey and [2] for a more recent one. For more recent work on temporal dynamics in remote sensing image data including change detection, see, for example, [3]–[5].

The method described here which is called iteratively reweighted multivariate alteration detection (IR-MAD) is a new extension to the previously published multivariate alteration detection (MAD) method [6]–[9] which, in turn, is based on the established multivariate statistical technique canonical correlation analysis (CCA), first described by Hotelling in 1936 [10]. Inspired by boosting methods often applied in data mining work [11] and by [12], iteratively reweighted MAD in a series of iterations places increasing focus on "difficult" observations; in a change detection setting, "difficult" observations are the ones whose change status over time is uncertain. This is done by calculating a measure of no change based on the sum of squared, standardized MAD variates in each iteration. This measure is then used as a weighting function for the calculation of the statistics used to calculate the MAD transformation in the next iteration. The idea in using such a scheme is to establish an increasingly better no-change background against which to detect change. Other types of robustification of the change detection method are briefly mentioned.

To prevent a change detection method from detecting uninteresting change due to noise or arbitrary spurious differences, this paper also describes the application of regularization (also known as penalization). Regularization in the form of smoothing of the (IR) CCA/MAD solution is described in some detail. The paper also mentions the exploitation of the affine transformation invariance of the MAD method as a regularizing measure, and a combination of the two types of regularization. Regularization may be important especially when change detection is applied to hyperspectral data.

Geometrical and other corrections required in order to carry out change studies are not dealt with here. For special problems with high spatial resolution, oblique viewing data, see, e.g., [13] and [14] on change detection in IKONOS data and [15] on change detection in QuickBird data.

Methods such as the ones described in this paper are well suited for inclusion in image processing packages and in geographical information systems (GIS).

Section II introduces multivariate change detection, Section III very briefly describes canonical correlation analysis, and Section IV defines the MAD transformation with subsections on both the suggested iterative re-weighting and regularization schemes. Section V gives three data examples with Landsat Thematic Mapper data (partly constructed data), SPOT High Resolution Visible data, and hyperspectral HyMap data. Section VI gives conclusion and dicusses future work. An Appendix gives more detail on selected aspects of canonical correlation analysis including regularization.

II. MULTIVARIATE CHANGE DETECTION

When we analyze changes in panchromatic image data with additive noise taken at different points in time, it is customary to calculate the difference between two images. The idea is, of course, that areas which exhibit no or small changes have zero or low absolute values and areas with large changes have large absolute values in the difference image. If we have two multivariate images with variables at a given location written as vectors (without loss of generality we assume that the expectation values $\mathrm{E}\{\boldsymbol{X}\} = \mathrm{E}\{\boldsymbol{Y}\} = \boldsymbol{0}$), $\boldsymbol{X} = [X_1 \cdots X_k]^T$ and $\boldsymbol{Y} = [Y_1 \cdots Y_k]^T$, where k is the number of spectral bands, then a simple spectral change detection transformation is the vector of band-wise differences also known as the change vector

$$\boldsymbol{X} - \boldsymbol{Y} = [X_1 - Y_1 \cdots X_k - Y_k]^T. \quad (1)$$

In general, simple differences make sense only if the data are normalized to a common zero and scale or calibrated over time. If our image data have (many) more than three spectral bands, it is difficult to visualize change in all bands simultaneously. To overcome this problem and to concentrate information on change, linear transformations of the image data that optimize some measure of change (also termed a design criterion) can be considered. A linear transformation that will maximize a measure of change in the simple multispectral difference image is one that maximizes deviations from no change, for instance the variance

$$\mathrm{V}\{v_1(X_1 - Y_1) + \cdots + v_k(X_k - Y_k)\}$$
$$= \mathrm{V}\{\boldsymbol{v}^T(\boldsymbol{X} - \boldsymbol{Y})\}. \quad (2)$$

Areas in the image data with high absolute values of $\boldsymbol{v}^T(\boldsymbol{X} - \boldsymbol{Y})$ are maximum change areas. A multiplication of vector $\boldsymbol{v}$ by a constant c will multiply the variance by c^2. Therefore, we must make a choice concerning $\boldsymbol{v}$. A natural choice is to request that $\boldsymbol{v}$ is a unit vector, $\boldsymbol{v}^T\boldsymbol{v} = 1$. Maximizing the variance in (2) under the constraint amounts to finding principal components of the simple difference images. Principal components analysis was developed by Hotelling in 1933 [16] based on a technique described by Pearson in 1901.

A more parameter rich measure of change that allows different coefficients for $\boldsymbol{X}$ and $\boldsymbol{Y}$ and different numbers of spectral bands in the two sets, p and q, respectively, $(p \leq q)$, are linear combinations

$$\boldsymbol{a}^T\boldsymbol{X} = a_1X_1 + \cdots + a_pX_p \quad (3)$$
$$\boldsymbol{b}^T\boldsymbol{Y} = b_1Y_1 + \cdots + b_qY_q \quad (4)$$

and the difference between them $\boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{Y}$. This measure in principle also accounts for situations where the spectral bands are not the same but cover different spectral regions, for instance if one set of data comes from the Landsat MultiSpectral Scanner (MSS) and the other set comes from the Landsat Thematic Mapper (TM) or from the SPOT High Resolution Visible (HRV) which may be valuable in historical change studies. In this case, one must be more cautious when interpreting the multivariate difference as multivariate change.

To find $\boldsymbol{a}$ and $\boldsymbol{b}$, [17] uses principal components (PC) analysis on $\boldsymbol{X}$ and $\boldsymbol{Y}$ considered as one concatenated vector variable; [18] applies PC analysis to simple difference images as described above. This approach requires normalized or calibrated data and results depend on the scale at which the individual variables are measured (for instance it depends on gain settings of a measuring device). Also, it forces the two sets of variables to have the same coefficients (with opposite signs), and it does not allow for the case where the two sets of images have different numbers of spectral bands.

Other change detection schemes based on simple difference images include factor analysis and maximum autocorrelation factor (MAF) analysis [19]–[21].

[12] deals with (iterated) PC analysis of the same variable at the two points in time and consider the second PC as a (marginal) change detector for that variable. [12] also introduces spatial measures such as inverse local variance weighting in statistics calculation and Markov random field modelling of the probability of change (versus no change).

Another approach is to define a set of $\boldsymbol{a}$ and $\boldsymbol{b}$ simultaneously in the fashion described below. Again, let us maximize the variance, this time $\mathrm{V}\{\boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{Y}\}$. A multiplication of $\boldsymbol{a}$ and $\boldsymbol{b}$ by a constant c will multiply the variance by c^2. Therefore, we must make choices concerning $\boldsymbol{a}$ and $\boldsymbol{b}$, and natural choices in this case are requesting unit variance of $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$, see Section III and the Appendix. The criterion then is maximize $\mathrm{V}\{\boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{Y}\}$ with $\mathrm{V}\{\boldsymbol{a}^T\boldsymbol{X}\} = \mathrm{V}\{\boldsymbol{b}^T\boldsymbol{Y}\} = 1$. With this choice, we have

$$\mathrm{V}\{\boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{Y}\}$$
$$= \mathrm{V}\{\boldsymbol{a}^T\boldsymbol{X}\} + \mathrm{V}\{\boldsymbol{b}^T\boldsymbol{Y}\} - 2\mathrm{Cov}\{\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}\}$$
$$= 2(1 - \mathrm{Corr}\{\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}\}). \quad (5)$$

We request that $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$ are positively correlated, see the next section on canonical correlation analysis. Therefore, determining the difference between linear combinations with maximum variance corresponds to determining linear combinations with minimum (non-negative) correlation. Determination of linear combinations with extreme correlations brings the theory of canonical correlation analysis to mind.

III. Canonical Correlation Analysis

Canonical correlation analysis investigates the relationship between two groups of several variables. It finds two sets of linear combinations of the original variables, one for each group. The first two linear combinations are the ones with the largest correlation. This correlation is called the first canonical correlation and the two linear combinations are called the first canonical variates. The second two linear combinations are the ones with the largest correlation subject to the condition that they are orthogonal to the first canonical variates. This correlation is called the second canonical correlation and the two linear combinations are called the second canonical variates. Higher order canonical correlations and canonical variates are defined similarly.

Since we are looking for canonical variates that are as similar as possible as measured by correlation, we request positive canonical correlations.

If we denote the variance-covariance matrix, also known as the dispersion (matrix), of the one set of variables (X) Σ_{11}, the dispersion of the other set of variables (Y) Σ_{22}, the covariance between them Σ_{12}, and the canonical correlation $\rho = \mathrm{Corr}\{a^T X, b^T Y\}$, we get (see the Appendix)

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a = \rho^2\Sigma_{11}a \qquad (6)$$
$$\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b = \rho^2\Sigma_{22}b \qquad (7)$$

or in terms of Rayleigh quotients

$$\rho^2 = \frac{a^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a}{a^T\Sigma_{11}a}$$
$$= \frac{b^T\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b}{b^T\Sigma_{22}b} \qquad (8)$$

i.e., we find the desired projections for X by considering the mutually orthogonal (also known as conjugate) eigenvectors $a_1, \ldots, a_p$ corresponding to the eigenvalues $\rho_1^2 \geq \cdots \geq \rho_p^2 \geq 0$ of $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ with respect to Σ_{11}. Similarly, we find the desired projections for Y by considering the conjugate eigenvectors $b_1, \ldots, b_p$ of $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ with respect to Σ_{22} corresponding to the same eigenvalues ρ_i^2.

This technique was first described in [10] and a treatment is given in most textbooks on multivariate statistics (good references are [22] and [23]).

Multiset canonical correlation analysis where we investigate the relationship between more than two groups of several variables first introduced in [24], [25] is described and applied to remote sensing data in [6] and [26]. Nonlinear (two- and multiset) canonical correlation analysis is dealt with in [27]–[30].

IV. MAD Transformation

Inspired by Sections II and III, we define the multivariate alteration detection (MAD) transformation as

$$\begin{bmatrix} X \\ Y \end{bmatrix} \rightarrow \begin{bmatrix} a_p^T X - b_p^T Y \\ \vdots \\ a_1^T X - b_1^T Y \end{bmatrix} \qquad (9)$$

where a_i and b_i are the defining coefficients from a standard canonical correlation analysis. To maximize variance in (5),

we must minimize ρ; therefore, we have reversed the order of the differences between the canonical variates in (9) so MAD variate 1 is the difference between the highest order canonical variates, MAD variate 2 is the difference between the second highest order canonical variates, etc.

The dispersion matrix of the MAD variates is

$$\mathrm{D}\{a^T X - b^T Y\} = 2(I - R) \qquad (10)$$

where I is the $p \times p$ unit matrix and R is a $p \times p$ matrix containing the ascendingly sorted canonical correlations on the diagonal and zeros off the diagonal so the MAD variates are orthogonal with variance

$$\sigma_{\mathrm{MAD}_i}^2 = 2(1 - \rho_{p-i+1}). \qquad (11)$$

The MAD transformation has the very important property that if we consider linear combinations of two sets X (of p variables) and Y (of q variables, $p \leq q$) that are positively correlated then the pth difference shows maximum variance among such variables. The $(p-j)$th difference shows maximum variance subject to the constraint that this difference is uncorrelated with the previous j ones. In this way, we sequentially extract uncorrelated difference images where each new image shows maximum difference (change) under the constraint of being uncorrelated with the previous ones. If $p < q$, then the projection of Y on the eigenvectors corresponding to the eigenvalues 0 will be independent of X. That part may be considered the extreme case of multivariate change detection.

As opposed to the principal components of simple differences, the MAD variates are invariant to affine transformations (including linear scaling), which means that they are sensitive to neither, for example, changes in gain settings and offset in a measuring device, nor to linear (affine) radiometric and atmospheric correction schemes.

Because the MAD variates are linear combinations of the measured variables, they will have approximately a Gaussian distribution because of the Central Limit Theorem, see, e.g., [31]. In addition, if there is no change at pixel j, then the ith MAD value, MAD_{ij}, has mean 0. Assuming also independence of the orthogonal MAD variates we may expect that the sum of the squared MAD variates for pixel j after standardization to unit variance approximately follows a χ^2 distribution with p degrees of freedom, i.e., approximately

$$T_j = \sum_{i=1}^{p} \left(\frac{\mathrm{MAD}_{ij}}{\sigma_{\mathrm{MAD}_i}}\right)^2 \in \chi^2(p). \qquad (12)$$

The standardization should ideally be done by means of the standard deviation of the no-change observations. This standard deviation can be estimated by means of expectation–maximization (EM) based methods for determining thresholds for differentiating between change and no change in the difference images, and for estimating the variance-covariance structure of the no-change observations [32]–[37]. Below, we use the standard deviation for all observations for simplicity. Provided that the proportion of changed pixels is small, this will have minimal effect.

Equation (12) can be used to assign labels "change" or "no-change" to each observation by means of percentiles in the

χ^2 distribution. We may choose to assign the label "change" to observations with χ^2 values greater than, say, the 99% percentile and similarly the label "no-change" to observations with χ^2 values smaller than, say, the 1% percentile. Since the MAD transformation is invariant to linear (affine) transformations these no-change observations are suitable for carrying out an automated normalization between the two points in time. This is described in detail in [38].

The spatial aspect introduced in change detection in [20] can be applied here also by postprocessing the MAD variates with the (change strength weighted) MAF transformation, see [8]. The spatial aspect is dealt with elegantly in a Markov random field setting in [33] and [34].

The main feature of the MAD method is the transformation from a space where the originally measured variables are ordered by wavelength into a feature space where the transformed, orthogonal variables are ordered by similarity (as measured by linear correlation). This latter ordering is considered to be more relevant for change detection purposes. Differences between corresponding pairs of variables in this latter space, i.e., differences between the canonical variates, give us the orthogonal MAD variates which can be considered as generalized difference images well suited for change detection.

A. Iteratively Reweighted MAD, IR-MAD

Inspired by boosting methods often used in data mining [11] and by [12], the idea in iteratively reweighted (IR) MAD is simply in an iterative scheme to put high weights on observations that exhibit little change over time. This can be done in several ways. We start with the original MAD transformation, i.e., we assign the same weight (= 1) to all pixels. A natural choice is to weight pixel j in the next iteration by w_j, which is a measure of no change, namely the probability of finding a greater value of the χ^2 value in (12)

$$w_j = P\{T_j > t\} \simeq P\{\chi^2(p) > t\}. \tag{13}$$

This weight enters into the calculation of mean values, variances and covariances (N is the number of pixels)

$$\bar{X}_i = \frac{\sum_{j=1}^{N} w_j X_{ji}}{\sum_{j=1}^{N} w_j} \tag{14}$$

for the mean value of X_i, and

$$S_{k\ell} = \frac{\sum_{j=1}^{N} w_j (X_{jk} - \bar{X}_k)(X_{j\ell} - \bar{X}_\ell)}{(N-1)\sum_{j=1}^{N} w_j / N} \tag{15}$$

for the covariance between X_k and X_ℓ (if $k = \ell$, we get the variance of X_k).

Iterations are performed until the largest absolute change in the canonical correlations becomes smaller than some preset small value ϵ, e.g., 10^{-6}. This weighting scheme maps the weights applied to the interval $[0, 1]$ and avoids very high weights. Unlike boosting methods, weights from the early iterations are not used in this scheme, only the weights from the final iteration are used so the "committee" and voting scheme often involved in boosting are not used here.

Of course, other reweighting schemes for example leading to robust estimation [39], [40] of the variance–covariance structure of the data could be used. A limited number of tests indicates that the iterated scheme suggested in this section performs better than robust estimation; see also [41].

B. Regularized IR-MAD

If we have many (correlated) variables, the solutions to the coupled generalized eigenvalue problems in (6) and (7) may become unstable due to (near) singular variance–covariance matrices causing small changes in the data to lead to dramatically different solutions. A possible solution to such (near) singularity problems in hyperspectral data change detection may be regularization (also known as penalization) where, inspired by ridge regression described in [42], we add $\lambda_1 \boldsymbol{\Omega}_1$ to $\boldsymbol{\Sigma}_{11}$ and $\lambda_2 \boldsymbol{\Omega}_2$ to $\boldsymbol{\Sigma}_{22}$ in (6) and (7). λ_i are (small) non-negative numbers that can be chosen subjectively or estimated from the data, see the Appendix. This was first described in the CCA context with $\boldsymbol{\Omega}_i$ as the identity matrix in [43]. [44] penalizes high local variation using a second order derivative-type $\boldsymbol{\Omega}_i$. To obtain a continuous and differentiable second order derivative, [45] in a functional setting suggests a fourth order derivative-type $\boldsymbol{\Omega}_i$.

Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ here are the same type of data, we choose $\lambda_1 = \lambda_2 = \lambda$ and $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$. We choose $\boldsymbol{L}$ in $\boldsymbol{\Omega} = \boldsymbol{L}^T \boldsymbol{L}$ to penalize curvature of the elements in $\boldsymbol{a}$ and $\boldsymbol{b}$ considered as functions of wavelength. Choosing the usual discrete approximation to the second order derivative we penalize $\boldsymbol{La}$ and $\boldsymbol{Lb}$ where $\boldsymbol{L}$ is $(p-2) \times p$ with (typically $p = q$)

$$\boldsymbol{L} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix} \tag{16}$$

leading to $p \times p$

$$\boldsymbol{\Omega} = \boldsymbol{L}^T \boldsymbol{L}$$
$$= \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -2 & 5 & -4 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & \cdots & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \tag{17}$$

(which is penta-diagonal and rank $p - 2$); see also [36], [46]–[50], and the Appendix.

Alternatively, the invariance of the MAD variates to linear (affine) transformations of the original variables can be exploited. Possible (near) singularities may also be remedied by means of principal component analysis (PCA), maximum autocorrelation factor (MAF), projection pursuit (PP) analysis or other dimensionality reducing projections applied to the variables at the two points in time separately before doing canonical correlation and MAD analysis. This approach is used in [49] and [51].
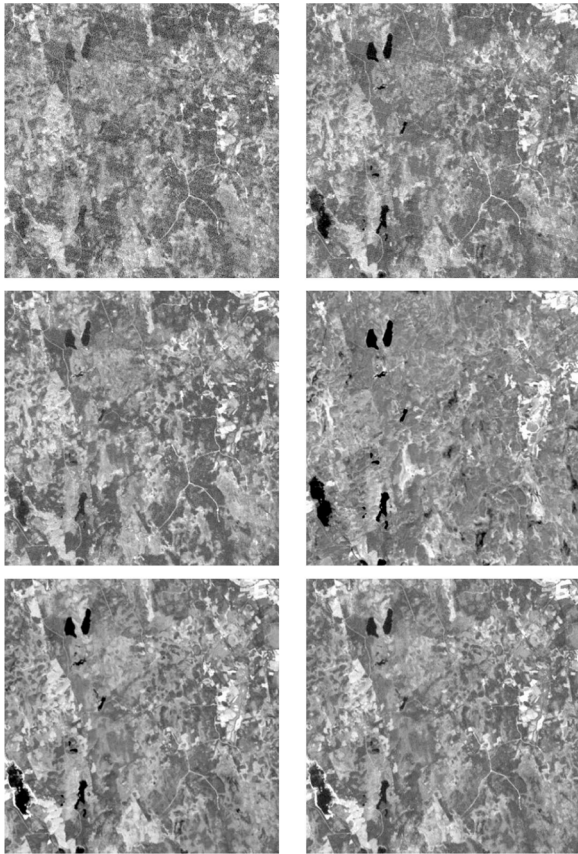
Fig. 1. Landsat TM data from June 6, 1986, covering a forested region in Northern Sweden, spectral bands 1, 2, 3, 4, 5, and 7 row-wise.
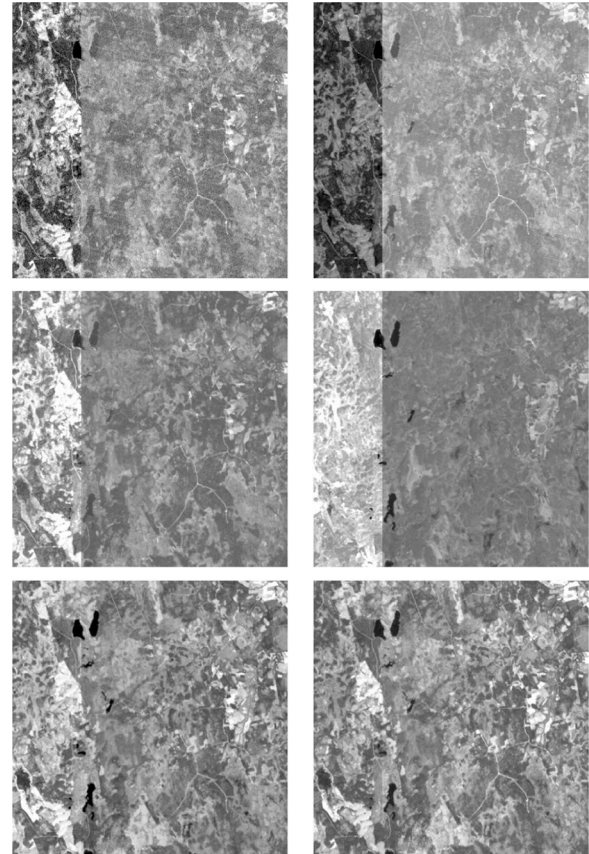

Fig. 2. Constructed Landsat TM data covering a forested region in Northern Sweden: the 512×128 leftmost part of the image consists of data from June 27, 1988, padded into the Landsat TM data from June 6, 1986, spectral bands 1, 2, 3, 4, 5, and 7 row-wise.

If regularization is needed or desired, one may use either the former, the latter or a combined scheme. The ordering of the projection variates in the dimensionality reducing regularization scheme is by some projection index (such as variance, autocorrelation, deviation from normality or other) rather than by wavelength. This ordering makes penalizing for example curvature un-natural. So the two regularization schemes do not readily combine.

Inspired by [52], we apply the dimensionality or feature reduction scheme above to adjacent, nonoverlapping groups of spectral bands. For example, we may replace bands 1, 2, and 3 with one projection, 4, 5, and 6 with another, etc. In this way, we reduce the dimensionality of the data (in the example by a factor of three) while retaining the main spectral features of the original data and their order. This preservation of order facilitates the application of further regularization by penalizing for example curvature as described in the former regularization scheme above.

If we use this combined regularization scheme, the general transformation invariance may be lost depending on the choice of dimensionality reduction scheme. If we choose the MAF transformation, we retain the invariance to any transformations that are linear (or affine) in the individual original variables.

V. Examples

The examples include a partly constructed case with Landsat Thematic Mapper (TM) data from a forested region in Northern Sweden, a case with SPOT High Resolution Visible (HRV) data covering an agricultural region in tropical Kenya, as well as a case with hyperspectral HyMap data from a small rural area in southeastern Germany.

All images in this paper are stretched linearly between mean and $\mp$ three standard deviations unless otherwise stated.

A. Partly Constructed Landsat TM Data, Northern Sweden

This case compares results from the original MAD method with those from the new iterated scheme where data at one point in time are constructed so that we know where change did not occur. Data at the first point in time are 512×512 25 m $\times$ 25 m Landsat Thematic Mapper (TM) spectral bands 1, 2, 3, 4, 5, and 7 from June 6, 1986, covering a forested region in Northern Sweden. Data at the second point in time are 512×128 Landsat TM (same bands) from June 27, 1988, covering the same region padded into the leftmost part of the 1986 image. Hence, by construction there is no change in the rightmost 512×384 part of the image. In this simple case, band-wise differences will give the desired zero change but as mentioned in general simple differences make sense only if the data are normalized to a common zero and scale or calibrated over time. Fig. 1 shows the measured 1986 data and Fig. 2 shows the partly constructed 1988 data.

Fig. 3 shows the original MAD variates and Fig. 4 shows the IR-MAD variates after 30 iterations (and convergence). Visual
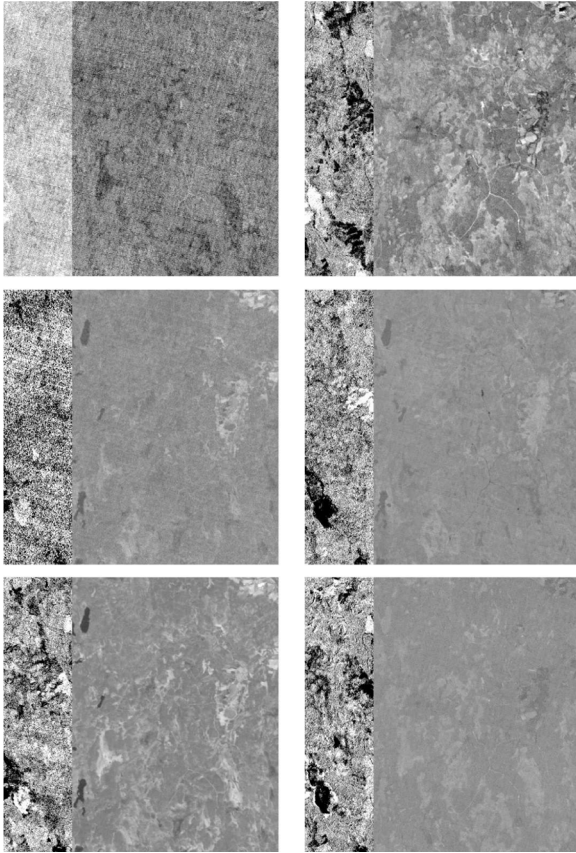
Fig. 3. Original MAD variates 1–6 row-wise. Very dark or very bright regions exhibit change, gray regions exhibit no change.
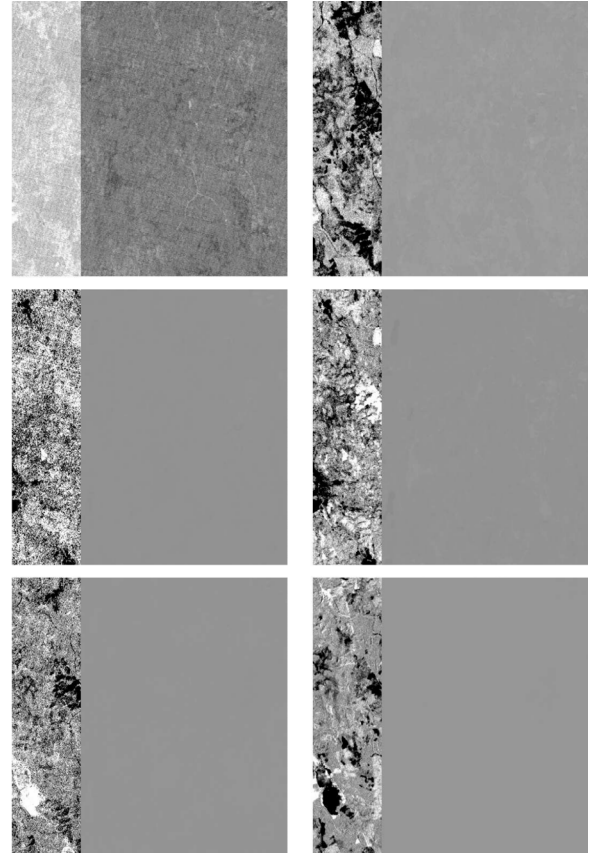

Fig. 4. IR-MAD variates 1–6 after 30 iterations row-wise. Very dark or very bright regions exhibit change, gray regions exhibit no change.

inspection shows that the iterated scheme does a much better job of finding no change in the rightmost 512×384 part of the image. Also, visual inspection indicates a less noisy appearance of the IR-MAD variates than of the original MAD variates in the leftmost 512×128 region of the image. This is supported by Table I which lists autocorrelations in that region between neighbouring pixels in the E-W, N-S, SW-NE, and SE-NW directions and their mean values, for both the original MAD and the IR-MAD variates.

Fig. 5 shows the development of the canonical correlations over the iterations. We see that the first iteration is most important and that most of the action takes place in the first 5–7 iterations.

Fig. 6 shows the (13) measure of no change for the original MAD variates (left) and for the IR-MAD variates (right) stretched linearly between 0 and 1. We see that after 30 iterations the weights assigned to the rightmost 512×384 no-change part of the image all remain close to one unlike the weights in the leftmost 512×128 potential change part. So the iterated scheme here provides a healthier background against which to detect change [Fig. 4 indicates that since practically only IR-MAD variate 1 has values different from zero in this case the sum in (13) should start with $i = 2$ rather than $i = 1$ to obtain a better no-change measure].

The mean, standard deviation, minimum, and maximum values for the original MAD variates for the rightmost 512×384 no-change region of the image are shown in

TABLE I

AUTOCORRELATIONS BETWEEN NEIGHBOURING PIXELS IN THE E-W, N-S, SW-NE, AND SE-NW DIRECTIONS AND THEIR MEAN VALUES, FOR BOTH THE ORIGINAL MAD AND THE IR-MAD VARIATES IN THE LEFTMOST 512×128 POTENTIAL CHANGE REGION OF FIGS. 3 AND 4

| | original MAD | | | | |
	E-W	N-S	SW-NE	SE-NW	mean
MAD1	0.58	0.55	0.35	0.48	0.49
MAD2	0.79	0.82	0.69	0.74	0.76
MAD3	0.52	0.61	0.34	0.43	0.48
MAD4	0.75	0.81	0.66	0.70	0.73
MAD5	0.72	0.74	0.58	0.65	0.67
MAD6	0.78	0.84	0.70	0.73	0.76
	IR-MAD				
	E-W	N-S	SW-NE	SE-NW	mean
MAD1	0.64	0.62	0.44	0.55	0.56
MAD2	0.88	0.90	0.81	0.84	0.86
MAD3	0.63	0.69	0.48	0.55	0.59
MAD4	0.86	0.87	0.78	0.81	0.83
MAD5	0.77	0.81	0.68	0.72	0.75
MAD6	0.92	0.94	0.88	0.90	0.91

Table II. The same quantities for the IR-MAD variates after 30 iterations for the rightmost 512×384 no-change region of the image are shown in Table III. All of these quantities (apart from the maximum of MAD variate 1) are closer to zero for the IR-MAD variates than for the original MAD variates indicating less change detected by the iterated MAD variates in the no-change region.
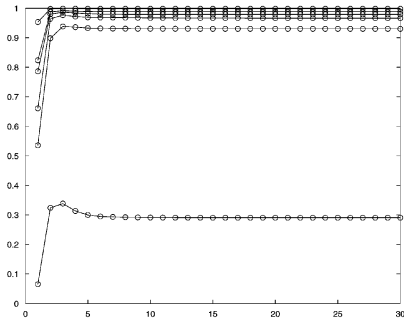
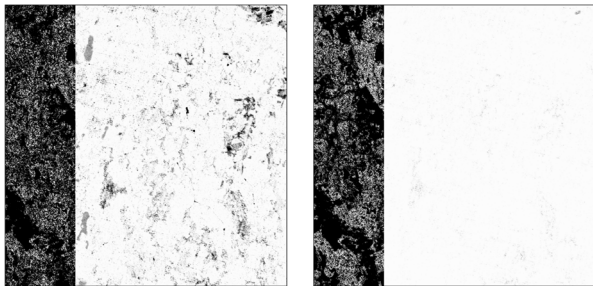Fig. 5. Canonical correlations over 30 iterations.

Fig. 6. Measure of no change for the original MAD variates (left) and for the IR-MAD variates after 30 iterations (right) both stretched linearly between 0 and 1. Bright areas are no change.

TABLE II
MEAN, STANDARD DEVIATION, MINIMUM, AND MAXIMUM VALUES
FOR THE ORIGINAL MAD VARIATES FOR THE RIGHTMOST 512×384
NO-CHANGE REGION OF THE IMAGE IN FIG. 3

	mean	stddev	min	max
MAD1	–0.59	0.87	–4.84	4.40
MAD2	0.029	0.61	–3.90	5.33
MAD3	–0.0093	0.26	–1.25	2.29
MAD4	–0.0080	0.10	–1.19	0.79
MAD5	–0.020	0.26	–1.17	2.00
MAD6	–0.0028	0.063	–0.37	0.48

Table IV shows mean, standard deviation, minimum, and maximum values for the sum of squared, standardized variates for the original and the IR-MAD transformations. For the entire image, we see that both the original MAD and IR-MAD method give the mean value we expect, namely six which is the number of degrees of freedom. Also, the greater standard deviation and range for the IR-MAD indicates a better discrimination between change and no change. For the no-change part of the image, Table IV shows that the IR-MAD method outperforms the original MAD method in this partly constructed no-change case.

Similar, but increasingly less pronounced, results (not shown) are obtained when the data at the second point in time are a 512×256 or a 512×384 (rather than a 512×128) 1988 scene padded into the 1986 scene so that the potential change region is two or three times larger.

To further illustrate the difference between the original and the iterated MAD variates, Fig. 7 shows the eigenvectors, i.e., the $\boldsymbol{a}$ s for the 1986 data and the $\boldsymbol{b}$ s for the 1988 data for the original MAD transformation. Fig. 8 shows these eigenvectors for

TABLE III
MEAN, STANDARD DEVIATION, MINIMUM, AND MAXIMUM VALUES
FOR THE IR-MAD VARIATES FOR THE RIGHTMOST 512×384
NO-CHANGE REGION OF THE IMAGES IN FIG. 4

	mean	stddev	min	max
IR-MAD1	–0.24	0.73	–4.47	5.85
IR-MAD2	–0.0054	0.095	–0.49	0.80
IR-MAD3	–0.000003	0.020	–0.081	0.18
IR-MAD4	–0.0017	0.033	–0.16	0.29
IR-MAD5	–0.00055	0.0092	–0.067	0.095
IR-MAD6	–0.00032	0.0036	–0.025	0.030

TABLE IV
MEAN, STANDARD DEVIATION, MINIMUM, AND MAXIMUM VALUES FOR THE
SUM OF SQUARED, STANDARDIZED ORIGINAL AND IR-MAD VARIATES FOR
THE ENTIRE IMAGE AND THE RIGHTMOST 512×384 NO-CHANGE
REGION OF THE IMAGES IN FIGS. 3 AND 4

	entire image			
	mean	stddev	min	max
MAD	6.00	16.2	0.023	466.0
IR-MAD	6.00	21.2	0.0079	634.2
	no-change region only			
	mean	stddev	min	max
MAD	1.36	1.54	0.0023	41.8
IR-MAD	0.45	0.408	0.0079	6.29

the IR-MAD transformation. Except for MAD6 corresponding to the leading or first canonical correlation we see that the eigenvectors are quite similar, IR-MAD4 plays the role of MAD5, and IR-MAD5 that of—MAD4. We also see that the eigenvectors for the iterated scheme are more symmetric than those of the original scheme, i.e., they are closer to the situation $\boldsymbol{a} = -\boldsymbol{b}$.

In [36], a study of bi-temporal Landsat TM data covering a semi-arid agricultural area in Hindustan, India, that uses the standard deviations for the no-change pixels only in the (13) no-change measure, shows that the IR-MAD variates clearly outperform the MAD variates in their ability to discriminate between change and no change. Also, the higher order IR-MAD variates are much less noisy than the higher order MAD variates (judged by visual inspection and measured again by the average spatial autocorrelations in the four main directions).

B. SPOT HRV Data, Kiambu District, Kenya

This case compares results from the original MAD method with those from the new iterated scheme where data at the first point in time are 512×512 20 m $\times$ 20 m SPOT High Resolution Visible (HRV) spectral bands 1, 2, and 3 from February 5, 1987 covering an agricultural region in Kenya. Data at the second point in time are 512×512 SPOT HRV (same bands) from February 12, 1989, covering the same geographical area. Fig. 9 shows the 1987 data and Fig. 10 shows the 1989 data.

Fig. 11 shows the original MAD variates and Fig. 12 shows the IR-MAD variates after 12 iterations (and convergence). Visual inspection shows that MAD variate 3 and to a lesser degree MAD variate 2 change substantially. It also indicates a less noisy appearance of IR-MAD variate 3 than of the original MAD variate 3. This is supported by Table V which lists autocorrelations between neighbouring pixels in the E-W, N-S, SW-NE, and SE-NW directions and their mean values, for both the original MAD and the IR-MAD variates.
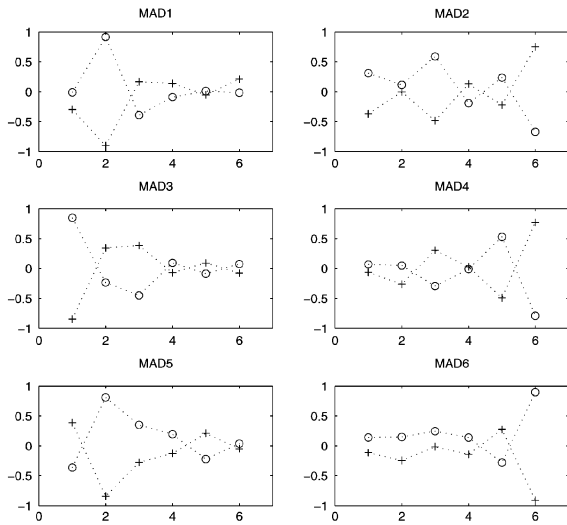
Fig. 7. Eigenvectors for the 1986 data (o) and the 1988 data (+) for the original MAD transformation.
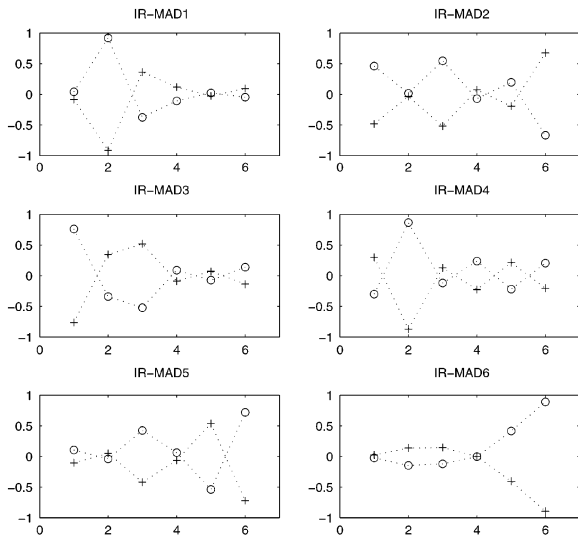

Fig. 8. Eigenvectors for the 1986 data (o) and the 1988 data (+) for the IR-MAD transformation.

Fig. 13 shows the development of the canonical correlations over the iterations. We see that the first iteration is most important and that most of the action takes place in the first 3–4 iterations.

Fig. 14 shows the (13) measure of no change for the original MAD variates (left) and for the IR-MAD variates (right). We see that after 12 iterations, the image of the measure seems less noisy (the mean autocorrelation between neighbouring pixels is 0.67 for original MAD and 0.73 for IR-MAD).

C. Hymap Data, Lake Waging-Taching, Germany

Two geometrically and atmospherically corrected HyMap [53] scenes with 126 spectral bands covering most of the wavelength region from 0.438 to 2.483 μm with 15–20 nm spacing acquired on June 30, 2003, at 8:43 UTC and August 4, 2003, at 10:23 UTC from a small area near Lake Waging-Taching, Bavaria, Germany, near the city of Salzburg, Austria, are used
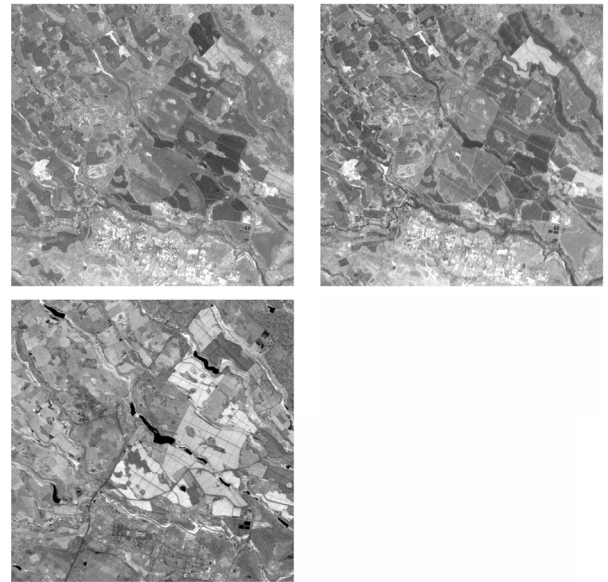

Fig. 9. SPOT HRV data from 5 February 1987 covering an agricultural region in Kenya, spectral bands 1, 2, and 3 row-wise.
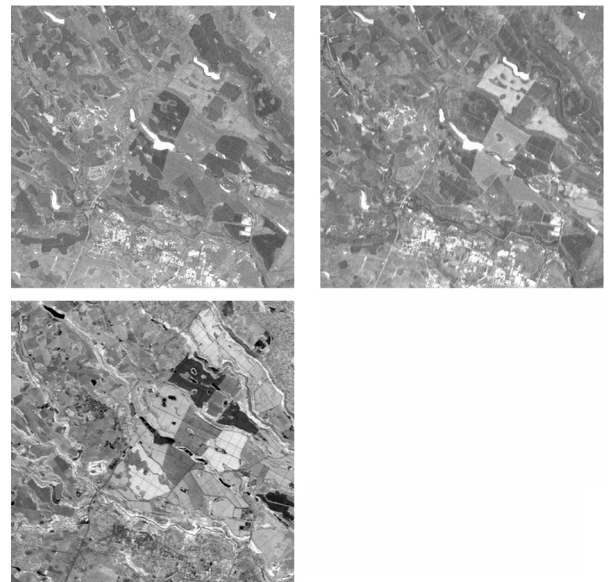

Fig. 10. SPOT HRV data from February 12, 1989, covering an agricultural region in Kenya, spectral bands 1, 2, and 3 row-wise.

to illustrate both the original MAD and the IR-MAD method including regularization with hyperspectral data. The image size is 400 by 270 (5 m $\times$ 5 m pixels). Fig. 15 shows HyMap bands 62, 40, and 19 on June 30, 2003, 8:43 UTC, and August 4, 2003, 10:23 UTC as RGB images.

Without some form of regularization the CCA and MAD processing, in this case, give IEEE standard "NaN" estimates of the lower order squared canonical correlations, which are clear signs of numerical (singularity) problems, i.e., the original MAD method does not work here.

The analyses carried out in [49] and [51] suggested a true dimensionality of around 40 for these data. Accordingly, in a regularization scheme which combines dimensionality reduction and curvature penalization, we choose 43 groups
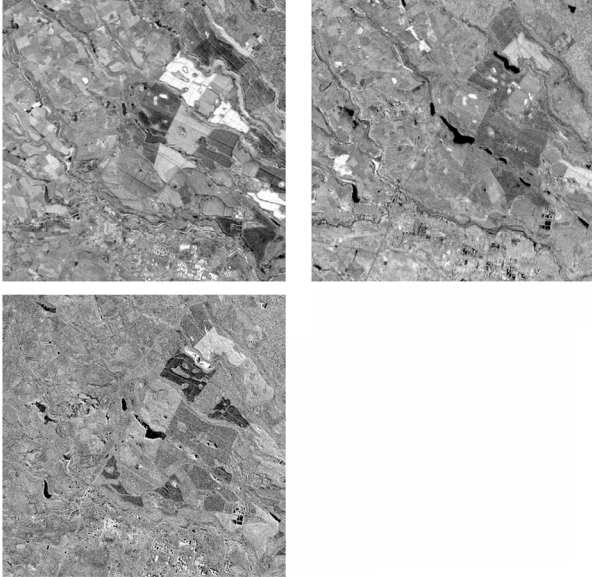
Fig. 11. Original MAD variates 1–3 row-wise. Very dark or very bright regions exhibit change, gray regions exhibit no change.
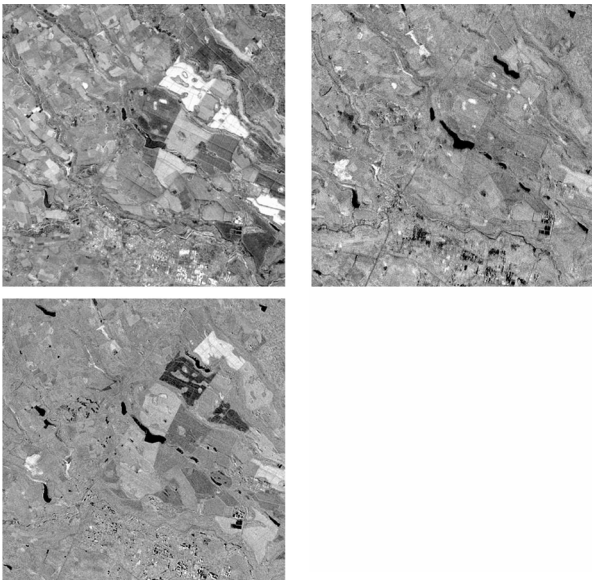

Fig. 12. IR-MAD variates 1–3 after 12 iterations row-wise. Very dark or very bright regions exhibit change, gray regions exhibit no change.

TABLE V

AUTOCORRELATIONS BETWEEN NEIGHBOURING PIXELS IN THE E-W, N-S, SW-NE, AND SE-NW DIRECTIONS AND THEIR MEAN VALUES, FOR BOTH THE ORIGINAL MAD AND THE IR-MAD VARIATES IN FIGS. 11 AND 12

	original MAD				
	E-W	N-S	SW-NE	SE-NW	mean
MAD1	0.89	0.87	0.80	0.84	0.85
MAD2	0.85	0.86	0.77	0.80	0.82
MAD3	0.64	0.64	0.48	0.52	0.57
	IR-MAD				
	E-W	N-S	SW-NE	SE-NW	mean
MAD1	0.88	0.87	0.79	0.83	0.84
MAD2	0.82	0.83	0.73	0.76	0.78
MAD3	0.81	0.82	0.69	0.75	0.77

with spectral bands $\{1, 2, 3\}, \{4, 5, 6\}, \ldots \{122, 123, 124\}$ and $\{125, 126\}$ for the dimensionality reducing projections (to
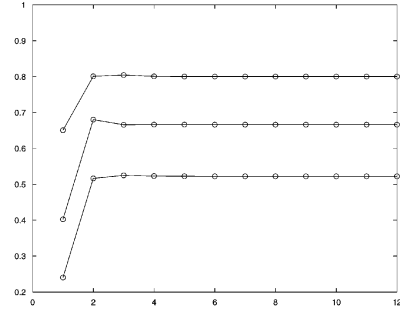

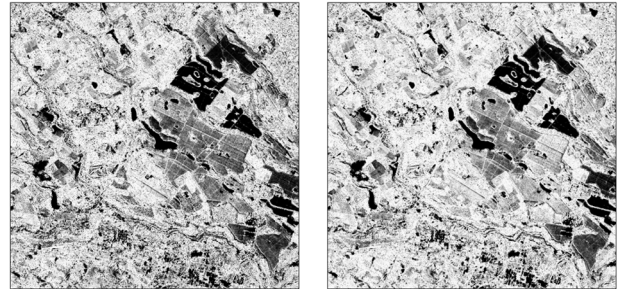Fig. 13. Canonical correlations over 12 iterations.

Fig. 14. Measure of no change for the original MAD variates (left) and for the IR-MAD variates after 12 iterations (right) both stretched linearly between 0 and 1. Bright areas exhibit no change.

avoid overlap between bands from HyMap's four detectors we use the following three two-bands-only groups: spectral bands $\{61, 62\}, \{93, 94\}$ and $\{125, 126\}$). Here, we use maximum spatial autocorrelation as the projection index. The obtained projection indices for all three (or two) obtainable projections are shown in Fig. 16 (the wavelength for a projection is chosen as the middle wavelength for the group if possible, if not the first wavelength is chosen). Only the projection corresponding to the highest projection index for each group (the top curve) is retained for further analysis.

Estimating λ for regularization of the projection variates simply by $\lambda = \text{trace } \Sigma_{11}/\text{trace } \Omega$ (see the Appendix) gives a λ value in the order of 10^{-1}. Tentative work on estimating λ from the leading pair of canonical variates by subsampling with fivefold cross validation [11] indicates λ values in the order of 10^{-2} to 10^{-4}. Using $\lambda = 10^{-3}$ gives quite wiggly weight functions indicating that even heavier regularization may be desirable. For illustration, we choose $\lambda = 10^{-1}$ here.

Fig. 17(a) shows the canonical correlations for the IR-MAD method with no curvature regularization over 100 iterations (this stabilizes the correlations to within less than 0.005). Fig. 17(b) shows the canonical correlations for the IR-MAD method with $\lambda = 0.1$ over the 22 iterations it takes to stabilize the correlations to within 0.001. On a Pentium III 1-GHz laptop with 1 Gbyte of memory this takes around 360 s corresponding to a little less than 16.5 s per iteration.

We see that without curvature regularization more iterations are needed compared to the situation with curvature regularization. In the latter case practical convergence seems to be obtained after maybe ten iterations here. Other runs (on 62 bands without the dimensionality reducing projection; not shown) indicate that with no or little regularization, large changes in some
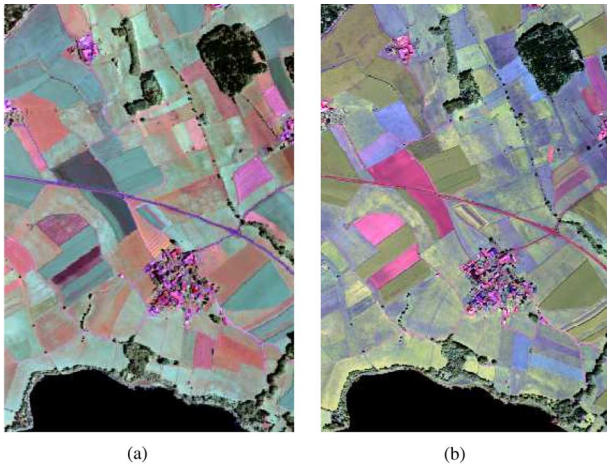
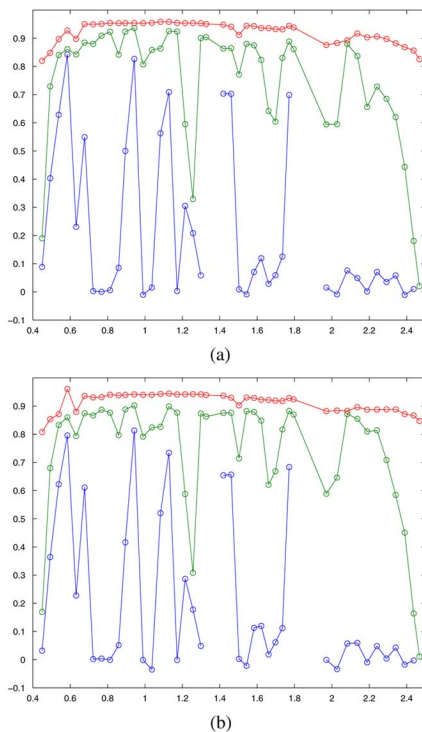Fig. 15. HyMap bands 62, 40, and 19 as RGB. (a) June 30, 2003, 8:43 UTC. (b) August 4, 2003, 10:23 UTC.

Fig. 16. Projection indices (spatial autocorrelations) for dimensionality reducing projections as functions of wavelength. (a) June 30, 2003. (b) August 4, 2003.

Fig. 17. Canonical correlations over iterations.

of the higher order canonical correlations occur up until around 150 iterations. With no or little regularization some precaution concerning convergence seems wise.

Fig. 18 shows the IR-MAD variates of the projections with no curvature regularization. Fig. 19 shows the IR-MAD variates of the projections with $\lambda = 0.1$. Fig. 20(a) and (b) shows the (12) χ^2 measure of change for no curvature regularization and for $\lambda = 0.1$, respectively, both stretched linearly between the 5% and 95% percentiles for the χ^2 distribution with 43 degrees of freedom, 28.96 and 59.30, respectively. Based on the visual inspection of Fig. 15 in which we see only three of the 126 original spectral bands, we see that several of the areas that seem
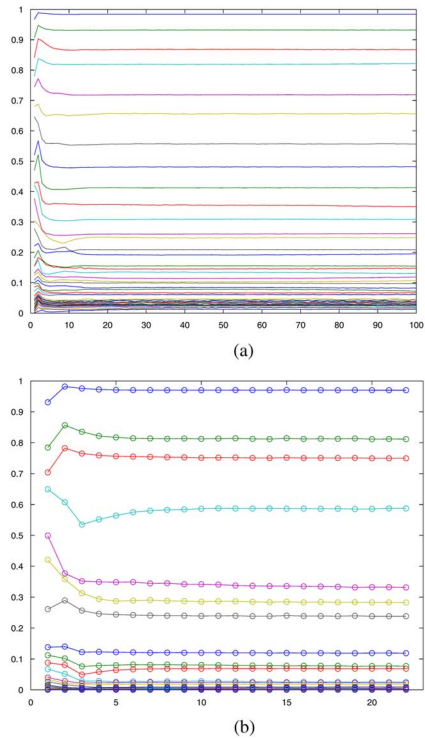
to change are more likely to be characterized as change regions in the regularized analysis. Also, the regularized versions of the change images appear less noisy.

ized IR-MAD variates 43, 42, and 41 as RGB. This plot shows several of the changes in the area between the two acquisition time points. Gray regions exhibit no change, regions with saturated colors (including black and white) exhibit change. Different colors represent different types of change. IR-MAD variates 43 and 42 (see Fig. 19 top left) are dominated by edge effects caused primarily by buildings, individual trees and groves combined with the difference in solar angles and possible geometric misregistrations between the acquisitions. To avoid these pronounced edge effects Fig. 21(b) shows the regularized IR-MAD variates 41, 40, and 39 as RGB. Again, gray regions exhibit no change, regions with saturated colors (including black and white) exhibit change.

Fig. 22(a) and (b) shows the eigenvectors (or weights) for the two leading canonical variates for no curvature regularization and for $\lambda = 0.1$, respectively. For the regularized situation, these curves are still somewhat wiggly which may hint that even stronger regularization could be applied.

In general, to interpret results from this and other types of change detection schemes it is recommended to perform simultaneous inspection and analysis of:

- the change images, here the MAD variates and the change/no-change measures;
- weight plots;
- spectra for selected pixels;
- results from clustering or classification of changes;
- mean spectra for selected groups or clusters of pixels;
- (per cluster) plots of correlations between original data and change variates.
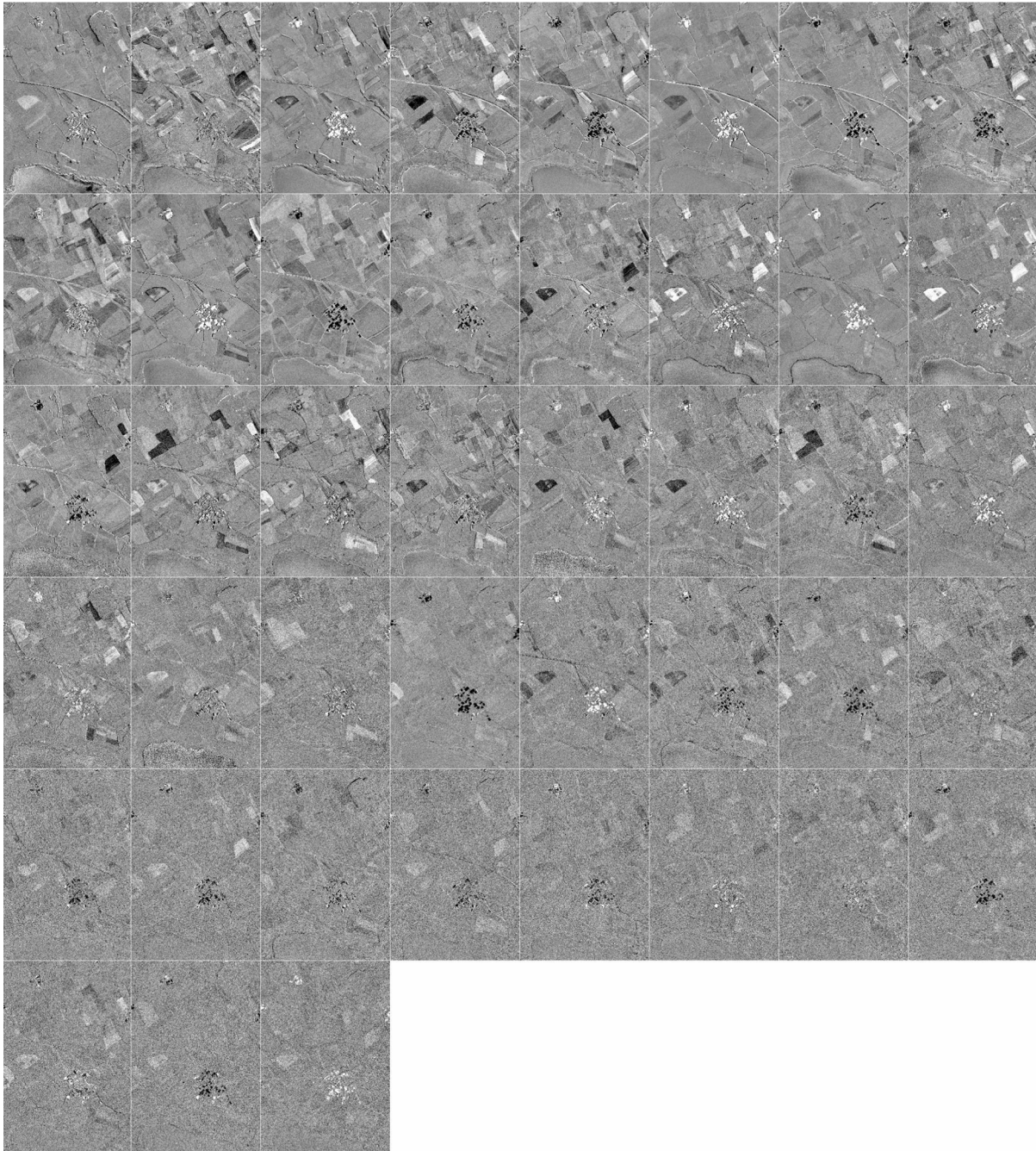
Fig. 18. IR-MAD variates 43-1 after 100 iterations row-wise, no curvature regularization. Very dark or very bright regions exhibit change, gray regions exhibit no change.

See also [37]. For space limitation reasons, only some of these recommendations are followed here.

VI. CONCLUSION AND FUTURE WORK

The iterated scheme described and applied to the partly constructed case with Landsat TM data clearly outperforms the original MAD scheme in terms of showing no change where no change should be. Also, in the region with potential change, higher autocorrelation, and, therefore, less noise in the change components are obtained with the iterated scheme.

For the SPOT HRV data, the autocorrelation for the last change component is improved drastically in the IR-MAD scheme with much less change between the MAD and IR-MAD schemes for the first two components.

In the example with hypervariate data, estimation of the MAD variates based on all 126 bands was not possible without some form of regularization.

For all three cases, we obtain a higher first order (also known as the leading) canonical correlation with the iterated scheme, indicating that a canonical correlation transformation to more similar variates is obtained. This shows that we do obtain a better background of no change against which to detect change.
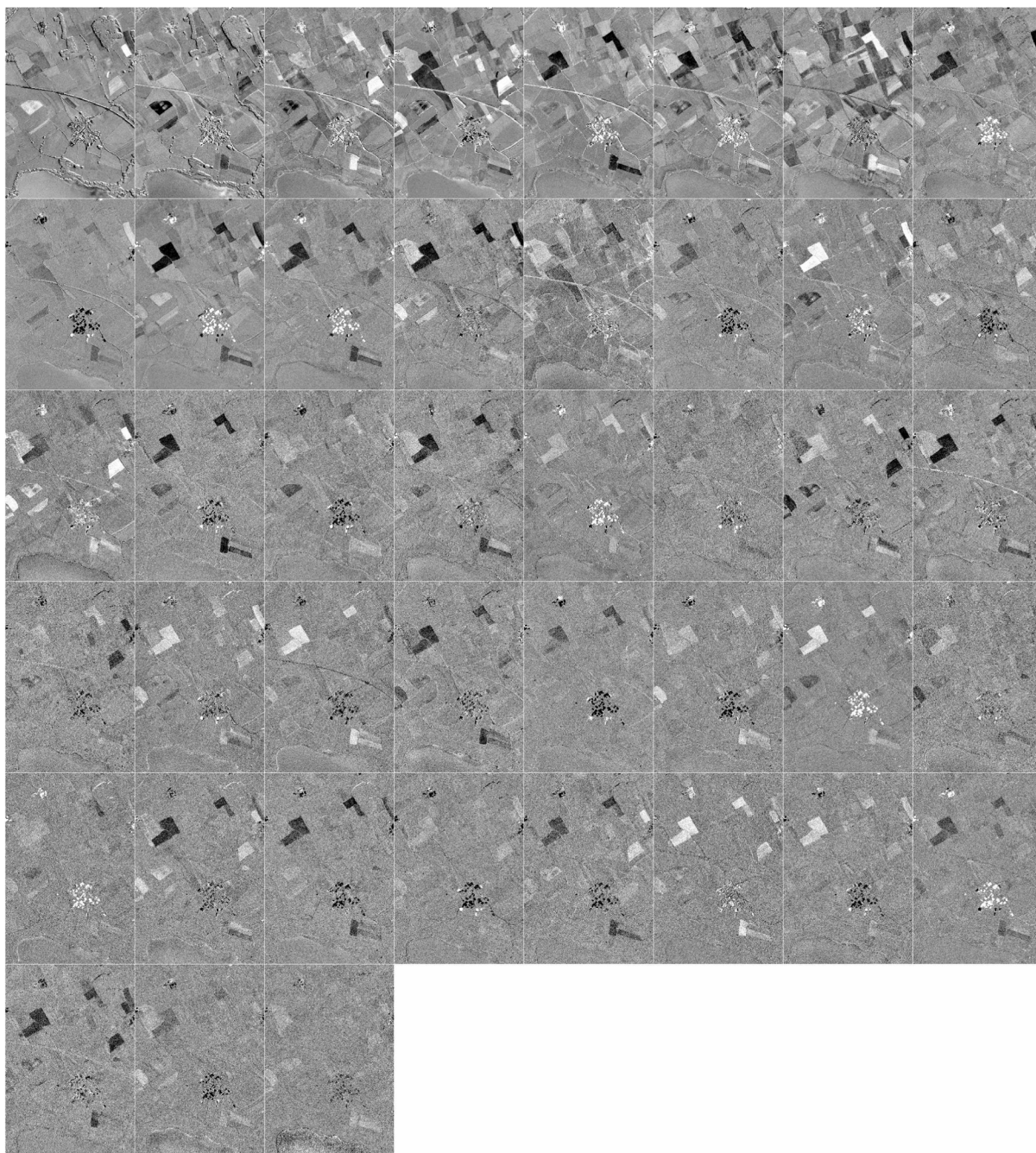
Fig. 19. IR-MAD variates 43-1 after 22 iterations row-wise, curvature regularization, $\lambda = 0.1$. Very dark or very bright regions exhibit change, gray regions exhibit no change.

In all three cases given, including the case with hypervariate data, the autocorrelation is higher and, hence, the noise lower in the no-change measure given in (13) for the IR-MAD scheme.

Since the methods described here are based on differences between canonical variates from a two-set canonical correlation analysis (CCA), they do not readily extend to a truly multitemporal setting where we have data from more than two time points. If such data were available, differences between canonical variates from a multiset CCA could be used as change variables. However, using the scheme suggested in this paper to pairs of bi-temporal data seems more viable to this author.

Multiset CCA could be applied to introduce a spatial element into the analysis by including spatially shifted versions of the bi-temporal data as new sets of variables.

Limited experience on the regularization scheme with hyperspectral data shows that more work could be done both on determining which and how many groups of spectral bands to choose in the dimensionality reducing projections, which projection index to choose, and on determining the regularization parameter λ and the matrix $\mathbf{\Omega}$.

A possible alternative to linear correlation as a measure of similarity between the transformed variables is mutual information, see, e.g., [54].
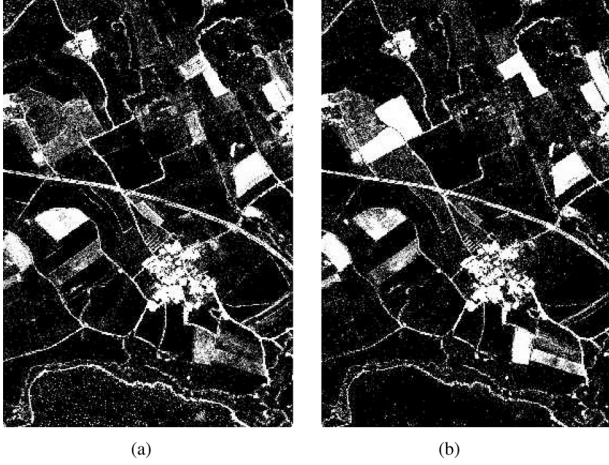
(a)

(b)

Fig. 20. χ^2 measure of change stretched linearly between 5% and 95% percentiles. Dark areas exhibit no change. (a) No curvature regularization. (b) Curvature regularization, $\lambda = 0.1$.
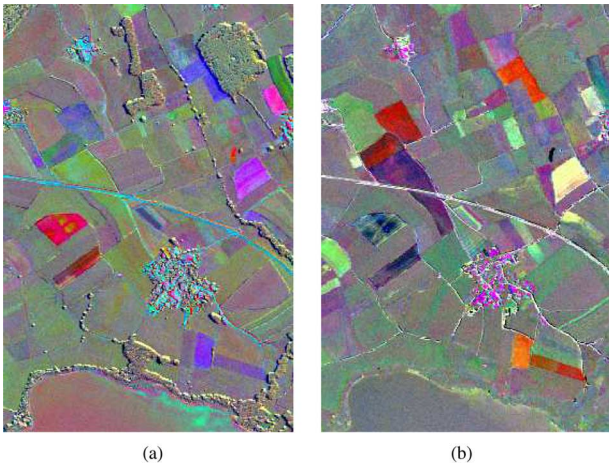

(a)

(b)

Fig. 21. IR-MAD variates as RGB, curvature regularization, $\lambda = 0.1$. Gray regions exhibit no change, regions with saturated colors (including black and white) exhibit change. (a) IR-MAD variates 43, 42, and 41. (b) IR-MAD variates 41, 40, and 39.
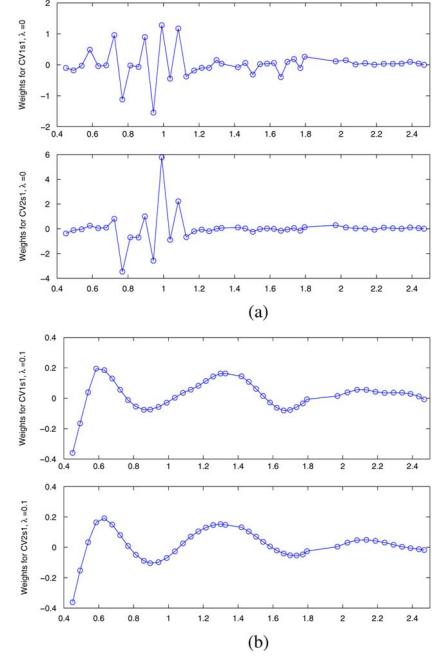

(a)

(b)

Fig. 22. Eigenvectors for the two leading canonical variates. (a) No curvature regularization. (b) Curvature regularization, $\lambda = 0.1$.

q positive definite dispersion Σ_{22}, $p \leq q$ for convenience) with maximal correlation

$$\rho = \mathrm{Corr}\{\boldsymbol{a}^T X, \boldsymbol{b}^T Y\}$$
$$= \frac{\mathrm{Cov}\{\boldsymbol{a}^T X, \boldsymbol{b}^T Y\}}{\sqrt{\mathrm{V}\{\boldsymbol{a}^T X\}\mathrm{V}\{\boldsymbol{b}^T Y\}}}. \qquad (18)$$

Without loss of generality, we assume that X and Y are zero mean, i.e., $\mathrm{E}\{X\} = \boldsymbol{0}$ and $\mathrm{E}\{Y\} = \boldsymbol{0}$. Since $\mathrm{Cov}\{\boldsymbol{a}^T X, \boldsymbol{b}^T Y\} = \boldsymbol{a}^T \Sigma_{12} \boldsymbol{b}$, where Σ_{12} is the covariance between X and Y, and the variances $\mathrm{V}\{\boldsymbol{a}^T X\} = \boldsymbol{a}^T \Sigma_{11} \boldsymbol{a}$ and $\mathrm{V}\{\boldsymbol{b}^T Y\} = \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}$, we get

$$\rho = \frac{\boldsymbol{a}^T \Sigma_{12} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}}. \qquad (19)$$

We request $\rho \geq 0$.

To maximize this correlation as a function of the coefficients in the linear combinations we find the partial derivatives of ρ with respect to $\boldsymbol{a}$ and $\boldsymbol{b}$

$$\frac{\partial \rho}{\partial \boldsymbol{a}} = \frac{\Sigma_{12} \boldsymbol{b} \sqrt{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}}{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}$$
$$- \frac{\boldsymbol{a}^T \Sigma_{12} \boldsymbol{b} \cdot 2 \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b} / (2\sqrt{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}})}{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}} \qquad (20)$$

$$\frac{\partial \rho}{\partial \boldsymbol{b}} = \frac{\Sigma_{21} \boldsymbol{a} \sqrt{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}}{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}$$
$$- \frac{\boldsymbol{a}^T \Sigma_{12} \boldsymbol{b} \cdot 2 \Sigma_{22} \boldsymbol{b} \cdot \boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} / (2\sqrt{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}})}{\boldsymbol{a}^T \Sigma_{11} \boldsymbol{a} \cdot \boldsymbol{b}^T \Sigma_{22} \boldsymbol{b}}. \qquad (21)$$

Another interesting future development of the MAD method lies in the functional setting described in [45].

Matlab code to carry out some of the analysis described is available from the author's homepage.

APPENDIX
CANONICAL CORRELATION ANALYSIS

Although canonical correlation analysis is described in most textbooks on multivariate statistics, see, e.g., [22] and [23], we give here a short derivation of some of the most important results including a regression analysis type interpretation and some regularization aspects.

A. Canonical Correlation

We are looking for linear combinations $\boldsymbol{a}^T X$ and $\boldsymbol{b}^T Y$ of two sets of multivariate observations X (p-dimensional with $p \times p$ positive definite dispersion Σ_{11}) and Y (q-dimensional with $q \times$

Setting $\partial\rho/\partial\boldsymbol{a} = \boldsymbol{0}$ $(p \times 1)$ and $\partial\rho/\partial\boldsymbol{b} = \boldsymbol{0}$ $(q \times 1)$, we obtain

$$\Sigma_{12}\boldsymbol{b} = \frac{\boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}}{\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a}}\Sigma_{11}\boldsymbol{a} \tag{22}$$

$$\Sigma_{21}\boldsymbol{a} = \frac{\boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}}{\boldsymbol{b}^T\Sigma_{22}\boldsymbol{b}}\Sigma_{22}\boldsymbol{b}. \tag{23}$$

Due to the symmetry of the two sets of variables, there is no reason not to choose $\mathrm{V}\{\boldsymbol{a}^T X\} = \mathrm{V}\{\boldsymbol{b}^T Y\}$, i.e., $\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} = \boldsymbol{b}^T\Sigma_{22}\boldsymbol{b}$. With this choice, we obtain from (22) and (23)

$$\Sigma_{12}\boldsymbol{b} = \rho\Sigma_{11}\boldsymbol{a} \tag{24}$$

$$\Sigma_{21}\boldsymbol{a} = \rho\Sigma_{22}\boldsymbol{b} \tag{25}$$

or

$$\begin{bmatrix} \boldsymbol{0} & \Sigma_{12} \\ \Sigma_{21} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \rho \begin{bmatrix} \Sigma_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \tag{26}$$

or

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = (\rho + 1) \begin{bmatrix} \Sigma_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \tag{27}$$

where $\rho = \boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}/\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} = \boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}/\boldsymbol{b}^T\Sigma_{22}\boldsymbol{b} = \boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}/\sqrt{\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} \cdot \boldsymbol{b}^T\Sigma_{22}\boldsymbol{b}}$. If $[\boldsymbol{a}^T \ \boldsymbol{b}^T]^T$ is an eigenvector with eigenvalue ρ, $[\boldsymbol{a}^T \ -\boldsymbol{b}^T]^T$ is an eigenvector with eigenvalue $-\rho$. We choose the positive eigenvalue.

We can choose the (equal) variances of $\boldsymbol{a}^T X$ and $\boldsymbol{b}^T Y$ freely. A natural choice is setting $\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} = \boldsymbol{b}^T\Sigma_{22}\boldsymbol{b} = 1$. In this case, $\rho = \boldsymbol{a}^T\Sigma_{12}\boldsymbol{b}$.

Alternatively, we could derive this generalized eigenvalue problem by requesting $\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} = \boldsymbol{b}^T\Sigma_{22}\boldsymbol{b} = 1$ to begin with and introduce these constraints into the optimization by means of Lagrange multipliers, see, e.g., [8] and [23].

With $\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a} = \boldsymbol{b}^T\Sigma_{22}\boldsymbol{b} = 1$, we get for (20) and (21)

$$\frac{\partial\rho}{\partial\boldsymbol{a}} = \Sigma_{12}\boldsymbol{b} - \rho\Sigma_{11}\boldsymbol{a} \tag{28}$$

$$\frac{\partial\rho}{\partial\boldsymbol{b}} = \Sigma_{21}\boldsymbol{a} - \rho\Sigma_{22}\boldsymbol{b}. \tag{29}$$

In this case, we get for the second order partial derivatives (also known as the Hessian)

$$\frac{\partial^2\rho}{\partial\boldsymbol{a}\partial\boldsymbol{a}^T} = -\rho\Sigma_{11} \tag{30}$$

$$\frac{\partial^2\rho}{\partial\boldsymbol{b}\partial\boldsymbol{b}^T} = -\rho\Sigma_{22}. \tag{31}$$

With $\rho > 0$, both Hessians are negative definite since both Σ_{11} and Σ_{22} are positive definite. This means that setting $\partial\rho/\partial\boldsymbol{a} = \boldsymbol{0}$ $(p \times 1)$ and $\partial\rho/\partial\boldsymbol{b} = \boldsymbol{0}$ $(q \times 1)$ do, indeed, lead to a maximum for ρ (as opposed to a minimum or a saddle point).

Another way of writing (26) is obtained by inserting $\boldsymbol{a}$ from (24) into (25), and by inserting $\boldsymbol{b}$ from (25) into (24). This gives the (more well-known) coupled eigenvalue problems for canonical correlation analysis

$$\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\boldsymbol{a} = \rho^2\Sigma_{11}\boldsymbol{a} \tag{32}$$

$$\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{b} = \rho^2\Sigma_{22}\boldsymbol{b}. \tag{33}$$

B. Interpretation of Canonical Variates

Consider a regression $\hat{X}$ of $X(p \times 1)$ based on $Y(q \times 1)$ and a regression $\hat{Y}$ of Y based on X respectively. This gives

$$\hat{X} = \Sigma_{12}\Sigma_{22}^{-1}Y \tag{34}$$

$$\hat{Y} = \Sigma_{21}\Sigma_{11}^{-1}X. \tag{35}$$

(In the univariate case, this latter equation reduces to the well known OLS regression expression $\hat{y} = s_{xy}/s_x^2 \cdot x$). For the dispersions, we get

$$\mathrm{D}\{\hat{X}\} = \Sigma_{12}\Sigma_{22}^{-1}\mathrm{D}\{Y\}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{36}$$

$$\mathrm{D}\{\hat{Y}\} = \Sigma_{21}\Sigma_{11}^{-1}\mathrm{D}\{X\}\Sigma_{11}^{-1}\Sigma_{12} = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{37}$$

and for the ratios of variances of linear combinations $\boldsymbol{a}^T\hat{X}$ and $\boldsymbol{b}^T\hat{Y}$ with linear combinations $\boldsymbol{a}^T X$ and $\boldsymbol{b}^T Y$, we get

$$\frac{\mathrm{V}\{\boldsymbol{a}^T\hat{X}\}}{\mathrm{V}\{\boldsymbol{a}^T X\}} = \frac{\boldsymbol{a}^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\boldsymbol{a}}{\boldsymbol{a}^T\Sigma_{11}\boldsymbol{a}} = \rho^2 \tag{38}$$

$$\frac{\mathrm{V}\{\boldsymbol{b}^T\hat{Y}\}}{\mathrm{V}\{\boldsymbol{b}^T Y\}} = \frac{\boldsymbol{b}^T\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{b}}{\boldsymbol{b}^T\Sigma_{22}\boldsymbol{b}} = \rho^2. \tag{39}$$

We see that the canonical variates can be interpreted as new variables that maximize the ratio of the variances between linear combinations of predicted values of one set of variables from the other set of variables and the same linear combinations of the actual values of the one set of variables.

We also see that unlike ordinary least squares regression analysis, canonical correlation analysis can be considered as a type of regression with several responses as well as several regressors with no distinction between responses and regressors.

C. Regularized Canonical Correlation

To remedy possible (near) singularity problems in Σ_{ii} which may occur in hyperspectral data, or to prevent the solutions $\boldsymbol{a}$ and $\boldsymbol{b}$ to the CCA/MAD problem to depend on noise or arbitrary spurious differences or simply to smooth the solutions (to facilitate interpretation), we may apply regularization (also known as penalization). This can be done by maximizing ρ in (18) subject to $\boldsymbol{a}^T(\Sigma_{11} + \lambda_1\Omega_1)\boldsymbol{a} = \boldsymbol{b}^T(\Sigma_{22} + \lambda_2\Omega_2)\boldsymbol{b} = 1$. This leads to

$$\begin{bmatrix} \boldsymbol{0} & \Sigma_{12} \\ \Sigma_{21} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}$$
$$= \rho \begin{bmatrix} \Sigma_{11} + \lambda_1\Omega_1 & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_{22} + \lambda_2\Omega_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}. \tag{40}$$

Here, the matrices $\Omega_i = \boldsymbol{L}_i^T\boldsymbol{L}_i$ penalize some characteristic $\boldsymbol{L}_1\boldsymbol{a}$ and $\boldsymbol{L}_2\boldsymbol{b}$ of the solutions and the $\lambda_i \geq 0$ determine the degree of penalization. Often, $\boldsymbol{L}_i$ is chosen to minimize the length (by setting $\boldsymbol{L}_1 = \boldsymbol{I}_p$ and $\boldsymbol{L}_2 = \boldsymbol{I}_q$, the $p \times p$ and $q \times q$ identity, respectively), slope or curvature of the solutions $\boldsymbol{a}$ and $\boldsymbol{b}$ (here considered as functions of wavelength) but also more complicated expressions that force the solutions to obey some ordinary differential equation can be used. To ensure the same influence of the regularization on all variables, it is customary to normalize them to (zero mean and) unit variance. Since all variables in our

case are digital numbers measured on the same scale, the normalization to unit variance is not performed here.

λ_i can be chosen subjectively or determined by a number of methods ranging from simply setting $\lambda_i = \text{trace } \boldsymbol{\Sigma}_{ii}/\text{trace } \boldsymbol{\Omega}_i$ to get the order of magnitude right, to subsampling with cross-validation [11] or L-curve estimation [47]. A fuller description of this subject is out of scope here. Other useful references are [42]–[46], [49], and [50].

ACKNOWLEDGMENT

REFERENCES

[1] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] P. Coppin, I. Jonckheere, K. Nackaerts, and B. Muys, "Digital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[3] L. Bruzzone and P. Smits, Eds., "Analysis of multi-temporal remote sensing images," in *MultiTemp Conf.*, Trento, Italy, Sep. 13–14, 2001.

[4] P. Smits and L. Bruzzone, Eds., "Analysis of multi-temporal remote sensing images," in *MultiTemp Conf.*, Ispra, Italy, Jul. 16–18, 2003.

[5] R. L. King and N. H. Younan, Eds., "Analysis of multi-temporal remote sensing images," in *MultiTemp Conf.*, Biloxi, MS, May 16–18, 2005.

[6] A. A. Nielsen, "Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data," Ph.D. dissertation, Inf. Math. Model., Tech. Univ. Denmark, Lyngby, (1994) [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?296

[7] A. A. Nielsen and K. Conradsen, "Multivariate alteration detection (mad) in multispectral, bi-temporal image data: a new approach to change detection studies," Tech. Rep. 1997-11, (1997) [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3092, Dept. Math. Model., Tech. Univ. Denmark, Lyngby.

[8] 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 Sens. Environ.*, vol. 64, pp. 1–19, (1998) [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?1220

[9] A. A. Nielsen, "Multi-channel remote sensing data and orthogonal transformations for change detection," in *Mach. Vis. Adv. Image Process. Remote Sens.*, I. Kanellopoulos, G. G. Wilkinson, and T. Moons, Eds., 1999 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?298

[10] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. XXVIII, pp. 321–377, 1936.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2001.

[12] R. Wiemker, A. Speck, D. Kulbach, H. Spitzer, and J. Beinlein, "Unsupervised robust change detection in multispectral imagery using spectral and spatial features," in *Proc. 2nd Int. Airborne Remote Sensing Conf. Exhibition*, ERIM, Ed., Copenhagen, Denmark, Jul. 7–10, 1997, vol. I, pp. 640–647.

[13] D. Haverkamp and R. Poulsen, *Change detection using IKONOS imagery*, 2003 [Online]. Available: http://www.spaceimaging.com

[14] I. Niemeyer and M. J. Canty, "Pixel-based and object-oriented change detection analysis using high-resolution imagery," presented at the 25th Symp. Safeguards and Nuclear Material Management, 2003 [Online]. Available: http://www.iniemeyer.de/publications/esarda03nie_cd.pdf

[15] I. Niemeyer, S. Nussbaum, and M. J. Canty, *Improvements of Object-Oriented Procedures for Detecting and Interpreting Changes Within Nuclear Facilities*, 2004 [Online]. Available: http://gmoss.jrc.cec.eu.int/workpackages/1/rma/20300_niemeyer_6m.pdf

[16] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Ed. Psych.*, vol. 24, pp. 417, 498–441, 520, 1933.

[17] T. Fung and E. LeDrew, "Application of principal components analysis to change detection," *Photogramm. Eng. Remote Sens.*, vol. 53, no. 12, pp. 1649–1658, 1987.

[18] P. Gong, "Change detection using principal component analysis and fuzzy set theory," *Can. J. Remote Sens.*, vol. 19, no. 1, pp. 22–29, 1993.

[19] P. Switzer and A. A. Green, "Min/Max Autocorrelation Factors for Multivariate Spatial Imagery," Tech. Rep. 6, Stanford Univ., Stanford, CA, 1984.

[20] P. Switzer and S. E. Ingebritsen, "Ordering of time-difference data from multispectral imagery," *Remote Sens. Environ.*, vol. 20, pp. 85–94, 1986.

[21] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.

[22] W. W. Cooley and P. R. Lohnes, *Multivariate Data Analysis*. New York: Wiley, 1971.

[23] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. New York: Wiley, 2003.

[24] J. D. Carroll, "A generalization of canonical correlation analysis to three or more sets of variables," in *Proc. 76th Convention of the American Psychological Association*, 1968, pp. 227–228.

[25] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, pp. 433–451, 1971.

[26] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multi-temporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.

[27] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 580–619, 1985.

[28] R. Coppi and S. Bolasca, Eds., *Multiway Data Analysis*. New York: Elsevier, 1989.

[29] K. Windfeld, "Application of Computer Intensive Data Analysis Methods to the Analysis of Digital Images and Spatial Data," Ph.D. dissertation, Inst. Math. Statist. Oper. Res., Tech. Univ. Denmark, Lyngby, 1992 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?1216

[30] K. B. Hilger, "Exploratory Analysis of Multivariate Data: Unsupervised Image Segmentation and Data Driven Linear and Nonlinear Decomposition," Ph.D. dissertation, Inf. Math. Model., Tech. Univ. Denmark, Lyngby, 2001 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?123

[31] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Belmont, CA: Wadsworth, 1995.

[32] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, no. 3, pp. 773–781, Mar. 1988.

[33] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 4, pp. 1171–1182, Apr. 2000.

[34] L. Bruzzone and D. F. Prieto, "An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 452–466, Mar. 2002.

[35] M. J. Canty and A. A. Nielsen, "Unsupervised classification of changes in multispectral satellite imagery," presented at the 11th SPIE Int. Symp. Remote Sensing, Maspalomas, Gran Canaria, Spain, Sep. 13–16, (2004) [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3245

[36] A. A. Nielsen and M. J. Canty, "Multi- and hyperspectral remote sensing change detection with generalized difference images by the IR-MAD method," presented at the MultiTemp Conf., Biloxi, MS, May 16–18, 2005.

[37] M. J. Canty and A. A. Nielsen, "Visualization and unsupervised classification of changes in multispectral satellite imagery," *Int. J. Remote Sens.* [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3889

[38] M. J. Canty, A. A. Nielsen, and M. Schmidt, "Automatic radiometric normalization of multitemporal satellite imagery," *Remote Sens. Environ.*, vol. 91, pp. 441–451, (2004) [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?2815

[39] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[40] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring, "Robust estimation of dispersion matrices and principal components," *J. Amer. Statist. Assoc.*, vol. 76, no. 374, pp. 354–362, 1981.

[41] W. Zhang, M. Liao, Y. Wang, L. Lu, and Y. Wang, "Robust approach to the MAD change detection method," in *Proc. 11th SPIE Int. Sympt. Remote Sensing X*, Maspalomas, Gran Canaria, Spain, Sep. 13–16, 2004, vol. 5574, pp. 184–193.

[42] A. E. Hoerl and R. W. Kennard, "Ridge regression. Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[43] H. D. Vinod, "Canonical ridge and econometrics of joint production," *J. Econometr.*, vol. 4, pp. 147–1662, 1976.

[44] B. Yu, I. M. Ostland, P. Gong, and R. Pu, "Penalized discriminant analysis of *in situ* hyperspectral data for conifer species recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 5, pp. 2569–2577, May 1999.

[45] J. Ramsay and B. Silverman, *Functional Data Analysis*. New York: Springer, 1997.

[46] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[47] P. C. Hansen, "Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion," SIAM, Philadelphia, PA, 1998.

[48] A. A. Nielsen and A. Müller, "Change detection by the MAD method in hyperspectral image data," in *Proc. 3rd EARSeL Workshop Imaging Spectroscopy*, (2003), pp. 115–116 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?2420

[49] A. A. Nielsen, "Regularisation in multi- and hyperspectral remote sensing change detection," presented at the 6th Geomatic Week Conf., Barcelona, Spain, Feb. 8–10, 2005 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3387

[50] A. A. Nielsen, "An iterative extension to the MAD transformation for change detection in multi- and hyperspectral remote sensing data," presented at the 4th EARSeL Workshop on Imaging Spectroscopy, Warsaw, Poland, Apr. 27–29, 2005 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3529

[51] A. A. Nielsen, A. Müller, and W. Dorigo, "Hyperspectral data, change detection and the MAD transformation," presented at the 12th Australasian Remote Sensing and Photogrammetry Association Conf., 2004 [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3176

[52] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.

[53] T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields, "The HyMap airborne hyperspectral sensor: The system, calibration, and performance," in *Proc. 1st EARSeL Workshop on Imaging Spectroscopy*, Zürich, Switzerland, Oct. 6–8, 1998, pp. 37–42 [Online]. Available: http://www.hyvista.com and http://www.intspec.com

[54] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, (2003) [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/itila

Allan Aasbjerg Nielsen received the M.Sc. degree from the Department of Electrophysics, Technical University of Denmark, Lyngby, in 1978, and the Ph.D. degree from Informatics and Mathematical Modelling (IMM), Technical University of Denmark, in 1994.

He is currently an Associate Professor with the IMM's section for Geoinformatics. He was with the Danish Defense Research Establishment from 1977 to 1978. He worked on energy conservation in housing with the Thermal Insulation Laboratory, Technical University of Denmark, from 1978 to 1985. Since 1985, he has been with the section for image analysis, IMM. Since then, he has worked on several national and international projects on the development, implementation, and application of statistical methods, and remote sensing in mineral exploration, mapping, geology, agriculture, environmental monitoring, oceanography, geodesy, and security funded by industry, the European Union, Danida (the Danish International Development Agency), and the Danish National Research Councils. His homepage can be found at URL http://www.imm.dtu.dk/~aa.