IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Anomaly Preserving $\ell_{2,\infty}$–Optimal Dimensionality Reduction over a Grassmann Manifold

Oleg Kuybeda, David Malah, and Meir Barzohar

CCIT Report #748
October 2009

Electronics
Computers
Communications

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT REPORT #748 October 2009

Anomaly Preserving $\ell_{2,\infty}$-Optimal Dimensionality Reduction over a Grassmann Manifold

Oleg Kuybeda, David Malah, and Meir Barzohar

Abstract

In this paper, we address the problem of redundancy reduction of high-dimensional noisy signals that may contain anomaly (rare) vectors, which we wish to preserve. Since anomaly data vectors contribute weakly to the ℓ_2-norm of the signal as compared to the noise, ℓ_2-based criteria are unsatisfactory for obtaining a good representation of these vectors. As a remedy, a new approach, named Min-Max-SVD (MX-SVD) was recently proposed for signal-subspace estimation by attempting to minimize the *maximum* of data-residual ℓ_2-norms, denoted as $\ell_{2,\infty}$ and designed to represent well both abundant and anomaly measurements. However, the MX-SVD algorithm is greedy and only approximately minimizes the proposed $\ell_{2,\infty}$-norm of the residuals. In this paper we develop an optimal algorithm for the minization of the $\ell_{2,\infty}$-norm of data misrepresentation residuals, which we call *Maximum Orthogonal complements Optimal Subspace Estimation* (MOOSE). The optimization is performed via a natural conjugate gradient learning approach carried out on the set of n dimensional subspaces in $\mathbb{R}^m$, $m > n$, which is a Grassmann manifold. The results of applying MOOSE, MX-SVD, and ℓ_2 - based approaches are demonstrated both on simulated and real hyperspectral data.

Index Terms

Signal-subspace rank, singular value decomposition (SVD), Min-Max-SVD (MX-SVD), Maximum Orthogonal-Complements Analysis (MOCA), Hyperspectral Signal Identification by Minimum Error (HySime), anomaly detection, dimensionality reduction, redundancy reduction, hyperspectral images, Grassmann manifold.

I. Introduction

Dimensionality reduction plays a key role in high-dimensional data analysis. In many sensor-array applications, meaningful signal structure belongs to a low-dimensional signal-subspace embedded in the high-dimensional space of the observed data vectors. There are many reasons that make dimensionality reduction of the observed data vectors crucial. For instance, dimensionality reduction allows improving SNR by eliminating dimensions that do not carry valuable signal information, but may contain noise that compromises the application performance; In applications such as anomaly detection and/or classification there is a problem related to high dimensional spaces due to so called *Hughes phenomenon* [1], according to which the performance of anomaly detection/classification algorithms significantly deteriorates when the number of training samples is severely limited for an accurate learning of the corresponding signal models; Dimensionality reduction allows reducing computational costs, as well as storage volumes. Numerous existing methods aim to estimate a low-dimensional signal-subspace that adequately reflects the meaningful signal structure. In this paper, we focus on applications that analyze data containing anomaly vectors in which the estimated signal-subspace should contain (preserve) anomaly vectors. The considered applications may require the estimated signal-subspace to be of a rank that is much lower than the observed dimensionality, and may be even lower than the physically meaningful signal structure. Such applications may be anomaly detection or classification, where Hughes phenomenon poses a serious problem for working in high-dimensional space, and in which the critical anomaly-related information should be retained even at the expense of the background information. Another example may be compression-related applications that may have similar background-anomaly related tradeoffs.

The commonly assumed observation model satisfies:

$$\mathbf{x}_i = \mathbf{A}\mathbf{s}_i + \mathbf{z}_i, \qquad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is the observed vector, $\mathbf{z}_i \in \mathbb{R}^p$ is the data-acquisition or/and model noise; $\mathbf{s}_i \in \mathbb{R}^r$, and $\mathbf{A} \in \mathbb{R}^{p \times r}$ is a full-rank matrix with rank r, ($r \leq p$). An example of application employing this model is anomaly detection in hyperspectral images. Here, the columns of $\mathbf{A}$ are the pure materials spectra (endmembers) and $\mathbf{s}_i$ their corresponding abundances [21]. [1]

A number of approaches have been proposed in the literature (e.g., [17], [18], [19]) for signal-subspace

[1] Due to physical reasons, $\{\mathbf{s}_i\}$ are constrained to be non-negative. However, for the dimensionality reduction that merely deals with the determination of the column space of $\mathbf{A}$ and not with the exact determination of $\mathbf{A}$ and/or $\{\mathbf{s}_i\}$, the constraints on $\{\mathbf{s}_i\}$ may be omitted and the pure signal vectors may be regarded as just a set of vectors lying in the column space of $\mathbf{A}$ without any relevance to $\{\mathbf{s}_i\}$.

estimation under the assumption that $\mathbf{s}_i$ and $\mathbf{z}_i$ are independent, stationary, zero-mean and Gaussian. It was shown in [2] that for white noise $\mathbf{z}_i$, the classical principal components analysis (PCA) method for signal subspace estimation is optimal in the Maximum-Likelihood (ML) sense. It determines the signal subspace by minimizing the ℓ_2-norm of misrepresentation residuals belonging to the complementary subspace, which can be obtained via Singular Value Decomposition (SVD) of $\mathbf{X}$, which is a matrix of observed data vectors $\{\mathbf{x}_i\}$ ordered as its colums. The authors of [15], propose a new ℓ_2-based approach, named as HySime, designed to determine both the signal subspace and its rank in hyperspectral imagery. The method first estimates the signal and noise covariance matrices. Then, they use the assumption on the nonnegativity of $\{\mathbf{s}_i\}$ in order to estimate the signal subspace rank by finding the subset of eigenvalues that best represents, in the ℓ_2-sense, the mean value of the data set. The signal subspace is obtained by applying SVD on the noise-reduced covariance matrix of the data. Unfortunately, as we show in [3], the ℓ_2-based criterion is unsatisfactory for obtaining a reliable representation of the anomaly (rare) vectors, which typically contribute weakly to the ℓ_2-norm of the signal as compared to the noise. Nevertheless, the proper representation of rare vectors may be of high importance in denoising and dimensionality reduction applications that aim to preserve all the signal-related information, including rare vectors, within the estimated low-dimensional signal-subspace. For example, in a problem of redundancy reduction in hyperspectral images, rare endmembers that are present in just a few data pixels contribute weakly to the ℓ_2-norm of the signal. Therefore, their contribution to the signal-subspace cannot be reliably estimated using an ℓ_2-based criterion. As a remedy, we propose in [3] a novel approach, named *Maximum Orthogonal-Complements Algorithm (MOCA)*, which employs a so-called $\ell_{2,\infty}$ norm for both *signal-subspace and rank determination*, designed to represent well both abundant and rare measurements, irrespective of their frequentness in the data. Mathematically, the $\ell_{2,\infty}$-norm of a matrix $\mathbf{X}$ is defined as follows:

$$\|\mathbf{X}\|_{2,\infty} \triangleq \max_{i=1,\ldots,N} \|\mathbf{x}_i\|_2, \tag{2}$$

where $\mathbf{x}_i$ denote columns of $\mathbf{X}$. In words: $\|\mathbf{X}\|_{2,\infty}$ means the *maximum* of ℓ_2-norms of $\mathbf{X}$ columns.

When $\ell_{2,\infty}$-norm is applied to the misrepresentation residuals, it penalizes individual data-vector misrepresentations, which helps to represent well not only abundant-vectors, but also rare-vectors. In [4] we show that the $\ell_{2,\infty}$-norm can be efficiently used for the detection of anomalies as well. However, the algorithm developed in [3] for signal-subspace estimation, named *Min-Max-SVD* (MX-SVD), is greedy and only approximately minimizes the proposed $\ell_{2,\infty}$-norm of misrepresentation residuals. In this paper we propose a new algorithm that utilizes a natural conjugate gradient learning approach proposed in

[5] to minimize $\ell_{2,\infty}$-norm of the misrepresentation residuals, where the signal-subspace basis matrix is constrained to the Grassmann manifold defined as the set of all n dimensional subspaces in $\mathbb{R}^m$, $n \leq m$ [5]. Since $\ell_{2,\infty}$-norm of the misrepresentation residuals can be also referenced as the maximum orthogonal complement norm, we denote the proposed approach as Maximum of *Orthogonal complements Optimal Subspace Estimation* (MOOSE).

This paper is organized as follows: In Section II we provide a brief overview of MX-SVD, the greedy algorithm for signal-subspace determination, proposed in [3]. In Section III we develop the proposed MOOSE algorithm. The results of applying MOOSE, SVD and MX-SVD are demonstrated on simulated (Section IV). The results of applying MOOSE, SVD, MX-SVD and HySime are demonstrated on real hyperspectral data (Section V). Finally, in Section VI, we conclude this work.

II. Overview of MX-SVD

In this section we provide a short overview of Min-Max SVD (MX-SVD), the greedy algorithm for signal-subspace determination, proposed in [3], designed to estimate an anomaly-preserving signal-subspace. Ideally, according to [3], given the estimated signal-subspace rank, k, the anomaly-preserving signal-subspace $\hat{\mathcal{S}}_k$ should satisfy:

$$\begin{aligned} \hat{\mathcal{S}}_k \quad &= \quad \underset{\mathcal{L}}{\text{argmin}} \;\; \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{2,\infty}^2 \\ & \qquad \text{s.t.} \qquad \text{rank}\;\mathcal{L} = k, \end{aligned} \tag{3}$$

where $\mathcal{P}_{\mathcal{L}^\perp}$ denotes an orthogonal projection onto $\mathcal{L}^\perp$. The greedy technique for the minimization of (3), used in [3], is to constrain the sought $\hat{\mathcal{S}}_k$ basis to be of the following form:

$$\hat{\mathcal{S}}_k = \text{range}\;\;[\boldsymbol{\Psi}_{k-h}|\boldsymbol{\Omega}_h]\,, \tag{4}$$

where $\boldsymbol{\Omega}_h$ is a matrix composed of h columns selected from $\mathbf{X}$, and $\boldsymbol{\Psi}_{k-h}$ is a matrix with $k-h$ orthogonal columns, obtained via SVD of $\mathcal{P}_{\boldsymbol{\Omega}_h^\perp}\mathbf{X}$. The main idea of MX-SVD is to collect anomaly vectors into $\boldsymbol{\Omega}_h$ in order to directly represent the anomaly vectors subspace. Since anomaly vectors are not necessarily orthogonal to background vectors, the matrix $\boldsymbol{\Omega}_h$ also partially represents background vectors. The residual background vector contribution to the null-space of $\boldsymbol{\Omega}_h^\top$ is represented by principal vectors found by applying SVD on $\mathcal{P}_{\boldsymbol{\Omega}_h^\perp}\mathbf{X}$.

The determination of the basis vectors of $\hat{\mathcal{S}}_k$ in terms of $[\boldsymbol{\Psi}_{k-h}|\boldsymbol{\Omega}_h]$ is performed as follows: First, we initialize $[\mathbf{\Psi}_k|\mathbf{\Omega}_0]$, such that

$$\mathbf{\Psi}_k = [\mathbf{u}_1, \ldots, \mathbf{u}_k]\,;\ \mathbf{\Omega}_0 = [], \tag{5}$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are k principal left singular vectors of $\mathbf{X}$.

Then, a series of matrices $\{[\mathbf{\Psi}_{k-j}|\mathbf{\Omega}_j]\}_{j=0}^{k}$ is constructed such that

$$\mathbf{\Omega}_{i+1} = [\mathbf{\Omega}_i|\mathbf{x}_{\omega_i}] \tag{6}$$

$$\mathbf{\Psi}_{k-i-1} = \left[\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-i-1}\right], \tag{7}$$

where, for each $i = 0, \ldots, k-1$, ω_i is the index of a data vector $\mathbf{x}_{\omega_j}$ that has the maximal residual squared norm r_i:

$$\omega_i \triangleq \underset{n=1,\ldots,N}{\text{argmax}} \|\mathcal{P}_{[\mathbf{\Psi}_{k-i}|\mathbf{\Omega}_i]^\perp}\, \mathbf{x}_n\|, \tag{8}$$

$$r_i \triangleq \|\mathcal{P}_{[\mathbf{\Psi}_{k-i}|\mathbf{\Omega}_i]^\perp}\, \mathbf{x}_{\omega_i}\|^2, \tag{9}$$

and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-i-1}$ are $k-i-1$ principal left singular vectors of $\mathcal{P}_{\mathbf{\Omega}_{i+1}^\perp}\mathbf{X}$. Thus, the k columns of $[\mathbf{\Psi}_{k-j}|\mathbf{\Omega}_j]$, for each $j = 0, \ldots, k$, span k-dimensional subspaces, respectively. Each subspace is spanned by a number of data vectors collected in the matrix $\mathbf{\Omega}_j$ and by SVD-based vectors that best represent (in ℓ_2 sense) the data residuals in the null-space of $\mathbf{\Omega}_j^\top$. Moreover, each subspace is characterized by it's maximum-norm misrepresentation residual r_j. The greedy signal-subspace estimation $\hat{\mathcal{S}}_k$ is selected as in (4), with

$$h = \underset{j=0,\ldots,k}{\text{argmin}}\ r_j. \tag{10}$$

This policy combines the ℓ_2-based minimization of background vector-residual norms with ℓ_∞-based minimization of anomaly vector residual norms, which produces a greedy estimate $\hat{\mathcal{S}}_k$ that approximately satisfies (3). A flowchart summarizing the MX-SVD process is shown in Fig. 1.

III. Minimizing $\ell_{2,\infty}$-norm on the Grassmann manifold

A. *Problem formulation*

Generally, the problem stated in (3) can be recast as

$$\hat{\mathcal{S}} = \underset{[\mathbf{W}]}{\text{argmin}}\ F([\mathbf{W}]), \tag{11}$$

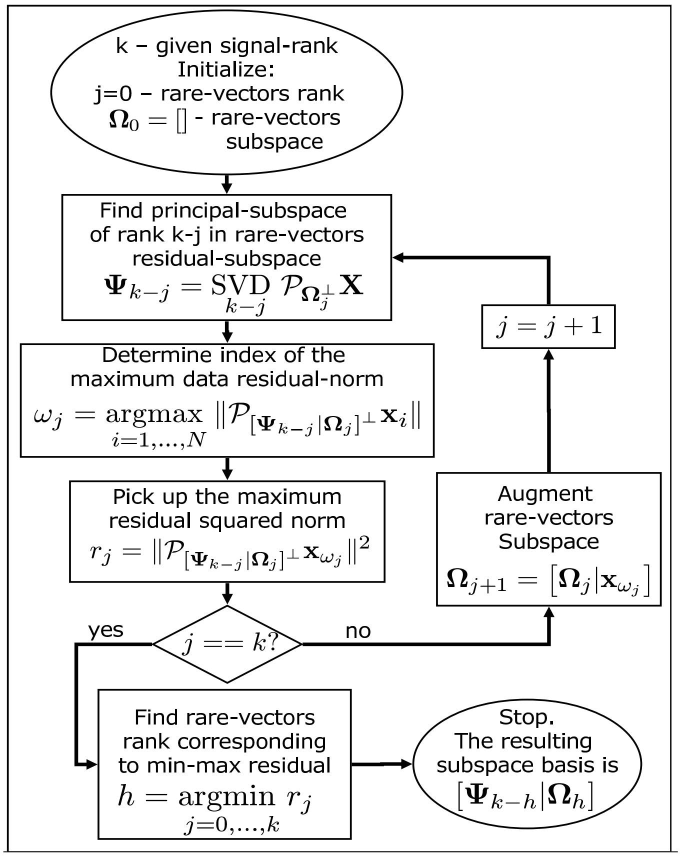
k – given signal-rank
Initialize:
j=0 – rare-vectors rank
$\mathbf{\Omega}_0 = []$ - rare-vectors subspace
Find principal-subspace of rank k-j in rare-vectors residual-subspace
$\mathbf{\Psi}_{k-j} = \underset{k-j}{\mathrm{SVD}}\ \mathcal{P}_{\mathbf{\Omega}_j^\perp}\mathbf{X}$
$j = j + 1$
Determine index of the maximum data residual-norm
$\omega_j = \underset{i=1,\ldots,N}{\mathrm{argmax}}\ \|\mathcal{P}_{[\mathbf{\Psi}_{k-j}|\mathbf{\Omega}_j]^\perp}\mathbf{x}_i\|$
Pick up the maximum residual squared norm
$r_j = \|\mathcal{P}_{[\mathbf{\Psi}_{k-j}|\mathbf{\Omega}_j]^\perp}\mathbf{x}_{\omega_j}\|^2$
Augment rare-vectors Subspace
$\mathbf{\Omega}_{j+1} = \left[\mathbf{\Omega}_j|\mathbf{x}_{\omega_j}\right]$
yes
$j == k?$
no
Find rare-vectors rank corresponding to min-max residual
$h = \underset{j=0,\ldots,k}{\mathrm{argmin}}\ r_j$
Stop. The resulting subspace basis is
$[\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$

Fig. 1. **MX-SVD flowchart.** For a given signal subspace rank value k, constructs a signal-subspace basis of the form $\hat{\mathcal{S}}_k = [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$, $h \in$ *integers* $[0, k]$, which approximately minimizes $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|^2_{2,\infty}$, where $\mathbf{\Omega}_h$ is responsible for representing anomaly-vectors and, partially, background vectors; $\mathbf{\Psi}_{k-h}$ complements $\mathbf{\Omega}_h$ to represent background vectors in the ℓ_2-sense.

where the objective function $F([\mathbf{W}])$ is defines as

$$F([\mathbf{W}]) \triangleq \|\mathbf{W}^\top\mathbf{X}\|^2_{2,\infty} \qquad (12)$$

and $[\mathbf{W}]$ is an equivalence class of all $p \times (p-k)$ orthogonal matrices whose columns span the same subspace in $\mathrm{I\!R}^p$ as $\mathbf{W}$. Here $[\mathbf{W}]$ represents the orthogonal complement subspace to the sought signal-subspace $\mathcal{S}_k$. The set of all n-dimensional subspaces in $\mathrm{I\!R}^m$, denoted by $G_{m,n}$, is called the Grassmann manifold [5]. The geometrical structure of the Grassmann manifold allows a continuous choice of subspaces, which is essential for constructing a local minimization procedure. Without loss of generality, by necessity, we must pick a representative of the equivalence class $[\mathbf{W}]$, say $\mathbf{W}$, in order to be able to work with $[\mathbf{W}]$ on the computer. Thus, by smoothly changing $\mathbf{W}$, such that $[\mathbf{W}] \in G_{p,p-k}$ we would be able to continuously move from one subspace to another and iteratively improve the objective function in a maner similar to well known unconstrained gradient-based algorithms such as steepest descent and conjugate gradient [23].

B. Grassmann manifold geometry

As stated in [5], the benefits of using gradient-based algorithms for the unconstrained minimization of an objective function can be carried over to a minimization constrained to the Grassmann manifold. The familiar operations employed by unconstrained minimization in the Euclidean space (plain space) such as computing gradients, performing line searches, etc., can be translated into their covariant versions on the Grassmann manifold (curved space).

In the following we briefly outline basic results from [5] used in this work for calculating gradients of an objective function and performing a line search along a search direction on the Grassmann manifold. Then, we develop a technique for minimizing $F([\mathbf{W}])$ of (12).

1) Gradient on Grassmann: The gradient of the objective function $F([\mathbf{W}])$ on the Grassmann manifold is defined to be a matrix $\nabla F \in T_{[\mathbf{W}]}$, where $T_{[\mathbf{W}]}$ is the tangent space at $[\mathbf{W}]$, such that for all $\mathbf{T} \in T_{[\mathbf{W}]}$, the following holds:

$$\langle F_{\mathbf{W}}, \mathbf{T} \rangle = \langle \nabla F, \mathbf{T} \rangle \,, \tag{13}$$

where $F_{\mathbf{W}}$ is the $p \times (p-k)$ matrix of partial derivatives of F with respect to the elements of $\mathbf{W}$; $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $p \times (p-k)$ - dimensional Euclidean space defined as

$$\langle \triangle_1, \triangle_2 \rangle \triangleq \mathrm{tr}(\triangle_1^\top \triangle_2). \tag{14}$$

In words, the relation in (13) states that the gradient of $F([\mathbf{W}])$ on the Grassmann manifold is the projection of $F_{\mathbf{W}}$ onto $T_{[\mathbf{W}]}$. Since $T_{[\mathbf{W}]}$ is the set of subspaces spanned by the columns of matrices of the form

$$\mathbf{T} = \mathbf{W}_\perp \mathbf{B}, \tag{15}$$

where $\mathbf{B}$ are arbitrary $k \times k$ matrices and $\mathbf{W}_\perp$ is a $p \times k$ orthogonal matrix satisfying

$$\mathbf{W}\mathbf{W}^\top + \mathbf{W}_\perp \mathbf{W}_\perp^\top = \mathbf{I}, \tag{16}$$

one obtains

$$\nabla F = F_{\mathbf{W}} - \mathbf{W}\mathbf{W}^\top F_{\mathbf{W}}. \tag{17}$$

A more rigorous treatment of these intuitive concepts is given in [5] where a solid foundation framework for the optimization algorithms involving orthogonality constraints is developed.

2) Line search: The line search in the Grassmann manifold is defined to be the minimization of $F([\mathbf{W}])$ along a geodesic, which is the curve of shortest length between two points in a manifold. By noticing that the geodesic equation is a second-order ODE, it follows from the local existence and uniqueness theorem that for any point $\mathbf{p}$ in a manifold and for any vector $\mathbf{v}$ in the tangent space at $\mathbf{p}$, there exists a unique geodesic curve passing through $\mathbf{p}$ in the direction $\mathbf{v}$ [6]. This observation makes the generalization of local optimization methods straightforward: given a descent direction $\mathbf{H} \in T_{[\mathbf{W}]}$ (for example, $\mathbf{H} = -\nabla F$), the objective function $F([\mathbf{W}])$ is minimized by the line search along the geodesic passing through $[\mathbf{W}]$ in the direction $\mathbf{H}$. An easy to compute formula for geodesics on the Grassmann manifold proposed in [5] reads as:

$$\mathbf{W}(t) = (\mathbf{W}\mathbf{V} \;\; \mathbf{U}) \begin{pmatrix} \cos{(t\mathbf{\Sigma})} \\ \sin{(t\mathbf{\Sigma})} \end{pmatrix} \mathbf{V}^\top, \tag{18}$$

where t is a geodesic curve traversing parameter and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the compact singular value decomposition (SVD) of $\mathbf{H}$. Compact SVD here means that the zero singular values are discarded along with the respective columns in $\mathbf{U}$ and $\mathbf{V}$, and the singular values are set in a decreasing order in $\mathbf{\Sigma}$. It can be easily verified that the diagonal elements of the matrix $t\mathbf{\Sigma}$ traverse Principal angles [10] between the column spaces $[\mathbf{W}(t)]$ and $[\mathbf{W}]$. Thus, for $t = 0$, one obtains the original subspace $[\mathbf{W}]$ that is rotated by the angles $t\mathbf{\Sigma}$ when t increases. Moreover, the geodesic distance between $[\mathbf{W}(t)]$ and $[\mathbf{W}]$ on the Grassmann manifold denoted by $d([\mathbf{W}(t)], [\mathbf{W}])$ satisfies [5]:

$$d([\mathbf{W}(t)], [\mathbf{W}]) = t\sqrt{\text{tr}(\mathbf{\Sigma}^2)}. \tag{19}$$

It should be noted that for large t values, the distance $d([\mathbf{W}(t)], [\mathbf{W}])$ is not the shortest one between $[\mathbf{W}]$ and $[\mathbf{W}(t)]$, since for large t, $[\mathbf{W}(t)]$ may complete one or more full circles in terms of the angles on the diagonal of $t\mathbf{\Sigma}$. However, it is still true that locally, for small t increments, $[\mathbf{W}(t)]$ is the shortest path on the Grassmann manifold connecting points on it. Moreover, the relation (19) implies that the rotation velocity, when one traverses the geodesic $[\mathbf{W}(t)]$ by changing t, equals to $\sqrt{\text{tr}(\mathbf{\Sigma}^2)}$ and, therefore, may change from iteration to iteration. In order to make it constant during the line search for all iterations, the matrix $\mathbf{\Sigma}$ is normalized:

$$\tilde{\mathbf{\Sigma}} \triangleq \mathbf{\Sigma} \big/ \sqrt{\text{tr}(\mathbf{\Sigma}^2)}. \tag{20}$$

Now, the line search is performed by looking for t that corresponds to a "significant reduction" of the objective function along a geodesic $[\mathbf{W}(t)]$. The notion of "a significant reduction" means that, on one hand, t should be low enough to ensure reduction of the objective function value; on the other hand, the search step t should be large enough for fast algorithm convergence. For this purpose, we use the Backtracking-Armijo linesearch method [23], [24] summarized in Algorithm 1.

Algorithm 1 *Backtracking-Armijo line search.*

```
    Given a geodesic $[\mathbf{W}(t)]$ in a descending direction $\mathbf{H}$, $\alpha \in (0, 0.5)$, $\beta > 1$, $t := t_0$
    Backtracking:
    while $(F([\mathbf{W}(t)]) > F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle)$, $t := t/\beta$
    Armijo:
    while $(F([\mathbf{W}(t)]) \leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle)$ and $(F([\mathbf{W}(\beta\, t)]) < F([\mathbf{W}]) + \alpha\, \beta\, t\, \langle \nabla F, \mathbf{H} \rangle)$, $t := \beta t$
```

In words, if the value of t is too large, it is iteratively decreased by dividing it by β in the Backtracking "while" stage, until the following condition holds:

$$F([\mathbf{W}(t)]) \leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle . \tag{21}$$

Since $\mathbf{H}$ is a descent direction and $\alpha < 1$, we have $\langle \nabla F, \mathbf{H} \rangle < 0$, so for small enough t, the following holds:

$$\begin{aligned} & F([\mathbf{W}(t)]) \approx F([\mathbf{W}]) + t\, \langle \nabla F, \mathbf{H} \rangle \leq \\ & \leq F([\mathbf{W}]) + \alpha\, t\, \langle \nabla F, \mathbf{H} \rangle \leq \\ & \leq F([\mathbf{W}]), \end{aligned} \tag{22}$$

which shows that the Backtracking "while" expression eventually terminates and that t is small enough to cause a decrease of the objective function value.

If the value of t is too small, it is iteratively increased by multiplying it by β in the Armijo "while" stage, until the condition (21) is concurrently satisfied with:

$$F([\mathbf{W}(\beta\, t)]) \geq F([\mathbf{W}]) + \alpha\, \beta\, t\, \langle \nabla F, \mathbf{H} \rangle . \tag{23}$$

In words, t is increased until it reaches a point in which it is still small enough to satisfy condition (21), but already large enough so that it is no longer satisfied in the next iteration, i.e., when βt replaces t (see (23)).

C. Minimization of $F([\mathbf{W}])$ on the Grassmann manifold.

In this subsection we develop a technique for solving (11) for $F([\mathbf{W}])$ of (12) on the Grassman manifold. A natural choice for the search direction is the negative gradient $\mathbf{H} = -\nabla F$ [7]. The calculation of ∇F involves the calculation of $F_{\mathbf{W}}$ (see (17)). For the calculation of $F_{\mathbf{W}}$ we consider here two cases: One case is when the maximum is obtained for only one data vector, while the other case is when the maximum is obtained for more than one data vector.

Case 1. If the maximum is obtained for only one vector at each $\mathbf{W}$ throughout the minimization, the calculation of $F_{\mathbf{W}}$ becomes straightforward:

$$F_{\mathbf{W}} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}, \tag{24}$$

where $\mathbf{x}_j$ is the vector for which $\max\limits_{i=1,\ldots,N} \|\mathbf{W}^\top \mathbf{x}_i\|_2$ is obtained.

Case 2. If the maximum is obtained for a set of indices J that contains more than one index, then the gradient direction $\hat{\mathbf{G}} = F_{\mathbf{W}} / \|F_{\mathbf{W}}\|_2$ is given by solving the following problem:

$$\begin{aligned} \hat{\mathbf{G}} \quad &= \quad \max_{\mathbf{G}} \min_{j \in J} \left\langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle \\ &\text{s.t.} \quad \left\langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle > 0 \;\; \forall j \in J \\ &\qquad\;\; \langle \mathbf{G}, \mathbf{G} \rangle = 1, \end{aligned} \tag{25}$$

with $\langle \cdot, \cdot \rangle$ being defined in (14). In words, it is a unit-norm matrix that maximizes the minimal projection norm onto gradients obtained individually for each $\mathbf{x}_j$, $j \in J$ (as in (24)). If the problem (25) is feasible, then the direction $-\hat{\mathbf{G}}$ is guaranteed to be a descent direction for all maximal residual norms $\|\mathbf{W}^\top \mathbf{x}_j\|$, $j \in J$, since all projections are constrained to be positive. Moreover, it is the steepest descent direction of the objective function $F([\mathbf{W}])$, because the descent rate of $F([\mathbf{W}])$ is determined by the lowest descent rate of the maximal residual norm $\|\mathbf{W}^\top \mathbf{x}_j\|$, for some $j \in J$, which is maximized (see the problem formulation in (25)). If the problem is infeasible, then $[\mathbf{W}]$ is a local minimum of the objective function $F([\mathbf{W}])$, since there is no search direction that concurrently minimizes all maximal residual norms. The problem (25) can be efficiently solved by Second-Order Cone Programming (SOCP) [23]. The norm of the derivative matrix $\|F_{\mathbf{W}}\|$ is given by

$$\|F_{\mathbf{W}}\| = \min_{j \in J} \left\langle \hat{\mathbf{G}}, \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W} \right\rangle, \tag{26}$$

I.e., it equals to the lowest descent rate of the the maximal residual norms, or equivalently, to the descent rate of $F([\mathbf{W}])$ in the direction $\hat{\mathbf{G}}$.

Practically, we have observed that in real data distributions the maximum is obtained for only one vector with probability close to one. Therefore, using (24) is good enough (practically) for obtaining a steep descent direction as we did in our simulations.

However, minimizing $F([\mathbf{W}])$ along the geodesic given by $-\nabla F$, may slow down the algorithm convergence due to an alternation of the competing maximum-norm data vectors from iteration to iteration. This phenomenon is also notoriously known as the zig-zag pattern pertaining to steepest descent methods [7]. In order to better cope with the complex nature of the cost function $F([\mathbf{W}])$, we propose to use the conjugate gradient method. According to this method, the conjugate search direction is a combination of the previous search direction and the new gradient

$$\mathbf{H}_{s+1} = -\nabla F_{s+1} + \gamma_s \tilde{\mathbf{H}}_s, \tag{27}$$

where s denotes the iteration index, $\tilde{\mathbf{H}}_s$ is the parallel translation of the previous search direction $\mathbf{H}_s$ from the point $[\mathbf{W}_s]$ to $[\mathbf{W}_{s+1}]$ by removing its normal component to the tangent space $\mathbf{T}_{\mathbf{W}s+1}$, as schematically shown in Fig. 2; and γ_s is obtained via Polak Ribiére conjugacy condition formula [5]

$$\gamma_s = \left\langle \nabla F_{s+1} - \tilde{\nabla} F_s, \nabla F_{s+1} \right\rangle \big/ \left\langle \nabla F_s, \nabla F_s \right\rangle, \tag{28}$$

where $\tilde{\nabla} F_s$ is the parallel translation of ∇F_s obtained in the same way as $\tilde{\mathbf{H}}_s$. The parallel translation is needed in order to keep all directions within the tangent space at each iteration. The formula for obtaining $\tilde{\nabla} F_s$ and $\tilde{\mathbf{H}}_s$ is [5]:

$$\begin{aligned} \tilde{\mathbf{H}}_s &= (-\mathbf{W}_s \mathbf{V} \sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U} \cos(t\tilde{\mathbf{\Sigma}}))\mathbf{\Sigma}\mathbf{V}^\top \\ \tilde{\nabla} F_s &= \nabla F_s - \\ &\quad (\mathbf{W}_s \mathbf{V} \sin(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}(\mathbf{I} - \cos(t\tilde{\mathbf{\Sigma}}))\mathbf{U}^\top \nabla F_s. \end{aligned} \tag{29}$$

The conjugate gradient construction offers a good compromise between convergence speed and computational complexity [9]. If the objective function is nondegenerate (locally quadratic), then the algorithm is guaranteed to converge quadratically in the Euclidean space [12]. The authors of [5] also show that in the Grassmann manifold, conjugate gradient algorithms also yield a quadratic convergence, i.e., for a manifold of dimension d, one has to perform a sequence of d steps to get to a distance within $O(\epsilon^2)$ from the solution. However, in our problem it is not guaranteed that the $\ell_{2,\infty}$-based cost function is locally

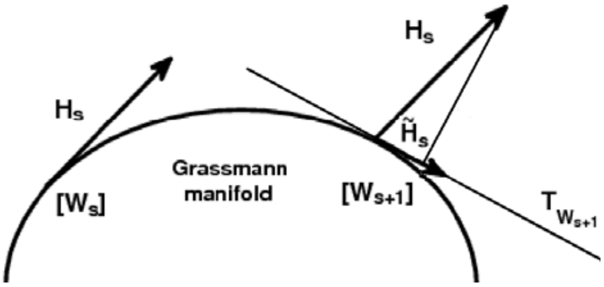

Fig. 2. **Parallel transport on Grassman manifold.**

quadratic. Therefore, there is no guarantee that the conjugate gradient descent procedure converges in d iterations. Fortunately, we have empirically found that the conjugate gradient descent method still significantly outperforms the gradient descent method in our problem. It is a common fact that conjugate gradient methods empirically still significantly outperform gradient descent methods even for nonconvex problems. In our case, a possible explanation to this may be as follows: The contribution of the previous search direction in each iteration, also helps the procedure to employ information that is carried in maximal norms obtained earlier (for possibly different data vectors). I.e., using the conjugate gradient direction helps to simultaneously minimize maximum-residual norms of vectors obtained in previous iterations. This helps to prevent the algorithm slow down due to an alternation of data-vectors corresponding to maximal-residual norms obtained from iteration to iteration.

As any local minimization of a non-convex objective function, the proposed algorithm is prone to getting trapped in a local minimum. Therefore, a proper initialization may be crucial for obtaining a good solution. Since MX-SVD finds a suboptimal solution using global principles, it provides a good initial point, which is close to the global minimum. Therefore, in our simulations we use the subspace obtained by MX-SVD as an initial point for the proposed approach.

The proposed approach for minimizing $F([\mathbf{W}])$ is summarized in Algorithm 2.

IV. Synthetic data simulation results

In this section we compare the results of applying SVD, MX-SVD and MOOSE to simulated examples in the presence of anomaly vectors. For this purpose the input data is constructed as follows:

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}, \tag{30}$$

with

$$\mathbf{Y} = \left[\sqrt{SNR_b}\,\mathbf{B}\mathbf{S}_b \mid \sqrt{SNR_a}\,\mathbf{A}\mathbf{S}_a\right], \tag{31}$$

Algorithm 2 *Conjugate gradient algorithm for minimizing $F([\mathbf{W}])$ on the Grassmann manifold.*

```
1 Given $\mathbf{W}_0$, such that $\mathbf{W}_0^\top \mathbf{W}_0 = \mathbf{I}_{p-k}$ and column space that coincides with the subspace obtained by MX-SVD, compute
  $F_{\mathbf{W}_0} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}_0,$ with j satisfying $\|\mathbf{W}_0^\top \mathbf{x}_j\|^2 = \|\mathbf{W}_0^\top \mathbf{X}\|_{2,\infty}^2$
  $\nabla F_0 = F_{\mathbf{W}0} - \mathbf{W}_0 \mathbf{W}_0^\top F_{\mathbf{W}0}$ and set $\mathbf{H}_0 = -\nabla F_0$
2 For $s = 0, 1, \ldots,$
    2.1 Obtain the compact decomposition of $\mathbf{H}_s$, $\mathbf{H}_s = \mathbf{U}\Sigma\mathbf{V}^\top$
    2.2 Normalize the principal angles $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma} \diagup \sqrt{\mathrm{tr}\mathbf{\Sigma}^2}$
    2.3 Perform Backtracking-Armijo line search (see Algorithm 1) along the geodesic
        $\mathbf{W}(t) = \mathbf{W}_s \mathbf{V}\mathrm{cos}(t\tilde{\mathbf{\Sigma}})\mathbf{V}^\top + \mathbf{U}\mathrm{sin}(t\tilde{\mathbf{\Sigma}})\mathbf{V}^\top$
    2.4 Update the subspace $\mathbf{W}_{s+1} = \mathbf{W}(t)$
    2.5 Parallel transport the tangent vectors $\mathbf{H}_s$ and $\nabla F_s$ to the point $[\mathbf{W}_{s+1}]$
        $\tilde{\mathbf{H}}_s = \left(-\mathbf{W}_s \mathbf{V}\mathrm{sin}(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}\mathrm{cos}(t\tilde{\mathbf{\Sigma}})\right) \mathbf{\Sigma}\mathbf{V}^\top$
        $\tilde{\nabla} F_s = \nabla F_s - \left(\mathbf{W}_s \mathbf{V}\mathrm{sin}(t\tilde{\mathbf{\Sigma}}) + \mathbf{U}(\mathbf{I} - \mathrm{cos}(t\tilde{\mathbf{\Sigma}})\right) \mathbf{U}^\top \nabla F_s$
    2.6 Compute the new gradients
        Euclidean: $F_{\mathbf{W}_{s+1}} = \mathbf{x}_j \mathbf{x}_j^\top \mathbf{W}_{s+1},$ with $j$ satisfying $\|\mathbf{W}_{s+1}^\top \mathbf{x}_j\|^2 = \|\mathbf{W}_{s+1}^\top \mathbf{X}\|_{2,\infty}^2$
        Grassmann: $\nabla F_{s+1} = F_{\mathbf{W}s+1} - \mathbf{W}_{s+1} \mathbf{W}_{s+1}^\top F_{\mathbf{W}s+1}$
    2.7 Compute the new search direction via Polak Ribiére conjugacy condition formula
        $\mathbf{H}_{s+1} = -\nabla F_{s+1} + \gamma_s \tilde{\mathbf{H}}_s$, where $\gamma_s = \left\langle \nabla F_{s+1} - \tilde{\nabla} F_s, \nabla F_{s+1} \right\rangle \diagup \langle \nabla F_s, \nabla F_s \rangle$
```

where $\mathbf{B}$ is a $p \times r_b$ matrix with orthogonal unit-norm columns spanning the background subspace; $\mathbf{A}$ is a $p \times r_a$ matrix with orthogonal unit-norm columns spanning the subspace of anomalies; $\mathbf{S}_b$ is a $r_b \times N_b$ matrix of background vector coefficients with columns drawn randomly from a Gaussian distribution with covariance matrix $\mathbf{C}_b = \mathbf{I} \diagup r_b$; $\mathbf{S}_a$ is a $r_a \times N_a$ matrix of anomaly vector coefficients with columns drawn randomly from a Gaussian distribution and *normalized to have unit-norm*; and $\mathbf{Z}$ is a $p \times (N_a + N_b)$ matrix containing white Gaussian noise with variance equal to $1/p$.

For SNR defined as

$$SNR \triangleq E\{\|\mathbf{y}\|^2\} \diagup E\{\|\mathbf{z}\|^2\}, \tag{32}$$

one can easily verify that background vectors have $SNR = SNR_b$, whereas the anomaly vectors have $SNR = SNR_a$. Moreover, due to the structure of the anomaly vector coefficient matrix $\mathbf{S}_a$, the norms of noise-free anomaly vectors are equal. This construction is designed to produce anomaly vectors that are equally significant.

Obviously, anomaly vectors are characterized by their low number compared to the number of background vectors, i.e., $N_a \ll N_b$. However, their number is allowed to be higher than the anomaly subspace dimension that they belong to, i.e., $N_a \geq r_a$. The extent of anomaly subspace population (loading) can be characterized by the loading ratio defined as follows:

$$R_a \triangleq N_a / r_a, \tag{33}$$

Thus, the minimal loading ratio $R_a = 1$ corresponds to the case where the number of anomalies is equal to the anomaly subspace rank. The larger the value of R_a is, the more anomaly vectors populate the anomaly subspace.

In our simulations we used the parameters shown in Table I. It is important to note that all parameters were selected to reflect a typical situation in hyperspectral images. Thus, SNR_a and SNR_b were selected to satisfy $SNR_a < SNR_b$ since the anomaly and the background subspaces in hyperspectral images are not orthogonal and, therefore, the anomaly vectors have weak orthogonal components to the subspace of background vectors.

TABLE I
MAXIMUM RESIDUAL-NORM SIMULATION PARAMETERS

p	r_b	r_a	N_b	N_a	SNR_b	SNR_a
100	5	5	10^5	10	100	10

In Fig. 3 one can see empirical pdfs of the maximum-residual norm $\|\mathbf{W}^\top \mathbf{X}\|^2_{2,\infty}$ obtained via a Monte-Carlo simulation, where $\mathbf{X}$ was generated 1000 times. As mentioned in [3], the estimated subspace by SVD may be skewed by noise in a way that completely misrepresents the anomaly vectors, since SVD uses ℓ_2 norm for penalizing the data misrepresentation, which is not sensitive to the anomaly-vector contributions. Hence, as clearly seen from the figure, the max-norm data residuals obtained by SVD (thick solid line) have high values which correspond to a poor representation of the anomaly vectors. It is also demonstrated in [3] that for $R_a = 1$ MX-SVD yields

$$\|\mathbf{W}^\top \mathbf{X}\|^2_{2,\infty} \approx \|\mathbf{W}^\top \mathbf{Z}\|^2_{2,\infty}. \tag{34}$$

In words, the empirical distribution of the maximum data residual norm $\|\mathbf{W}^\top \mathbf{X}\|^2_{2,\infty}$ for $R_a = 1$ is very close to the distribution of the maximum residual norm of noise $\|\mathbf{W}^\top \mathbf{Z}\|^2_{2,\infty}$, which has a limiting distribution known as the Gumbel distribution [22] (plotted in thin solid line in Fig. 3). However, as seen in that figure, for $R_a > 1$ (in this simulation $R_a = 2$), MX-SVD produces max-norm data residuals (whose pdf is plotted in dashed line) that are higher than the max-norm noise residuals. This happens since MX-SVD estimates the anomaly subspace by directly selecting r_b anomalous vectors from the data that contain noise, which skews the resulting subspace. The result is significantly improved by applying the optimal approach which produces max-norm data residuals (whose pdf is plotted in dot-dashed line) with values that are even lower than one would obtain from the Gumbel distribution.

The paradox of such a "super-efficiency" of the optimal approach is explained as follows: On one hand, the Gumbel distribution approximation is valid for max-norm realizations of data vectors drawn from Gaussian distribution. On the other hand, the max-norm data residuals obtained by MOOSE stem no longer from a Gaussian distribution, since they are minimized by MOOSE and, as a result, become lower than if the corresponding data vectors where randomly sampled from a Gaussian distribution.

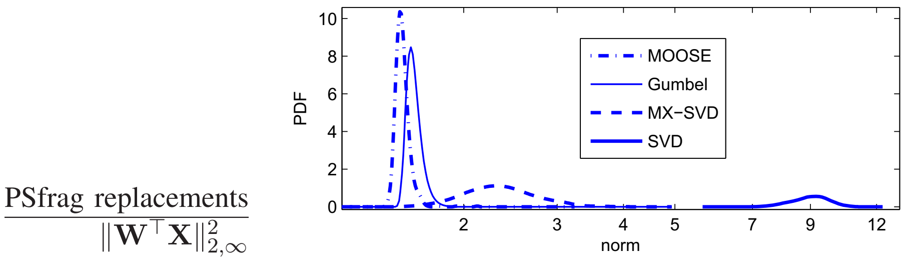

Fig. 3. **The pdfs of $\|\mathbf{W}^\top\mathbf{X}\|_{2,\infty}^2$ obtained via Monte-Carlo simulation.** The empirical pdfs of $\|\mathbf{W}^\top\mathbf{X}\|_{2,\infty}^2$ obtained by SVD (thick solid line), MX-SVD (dashed line), MOOSE (dot-dashed line) and the limiting Gumbel distribution approximating maximum residual norm of noise (thin solid line).

In Fig. 4 we compare SVD, MX-SVD and the proposed MOOSE algorithm in terms of subspace estimation error. The subspace error used here is defined to be the largest principal angle $\angle\{\hat{\mathcal{S}}, \mathcal{S}\}$ defined as follows [11]:

$$\angle\{\hat{\mathcal{S}}, \mathcal{S}\} = \max_{\mathbf{u}\in\hat{\mathcal{S}}} \min_{\mathbf{v}\in\mathcal{S}} \angle\{\mathbf{u}, \mathbf{v}\}, \quad \mathbf{u} \neq 0, \mathbf{v} \neq 0, \tag{35}$$

where $\hat{\mathcal{S}}$ and $\mathcal{S}$ denote the estimated subspace and the original subspace used for the data generation, respectively. In our simulations, for each R_a value $\mathbf{X}$ was generated 50 times. The considered R_a values were sampled logarithmically in $[1, 40]$ as shown in Fig. 4. For each R_a value we plot the mean of the subspace estimation error values obtained by SVD (line with star marks), MX-SVD (line with circle marks) and the proposed approach (line with diamond marks). As clearly seen from the figure, the proposed approach corresponds to the lowest mean subspace estimation error for all R_a values. The MX-SVD and the proposed approach perform much better than SVD for a wide range of R_a values. For R_a values high enough SVD manages to catch up with the other two $\ell_{2,\infty}$-norm based approaches, since then the anomalies become significant in terms of the ℓ_2-norm.

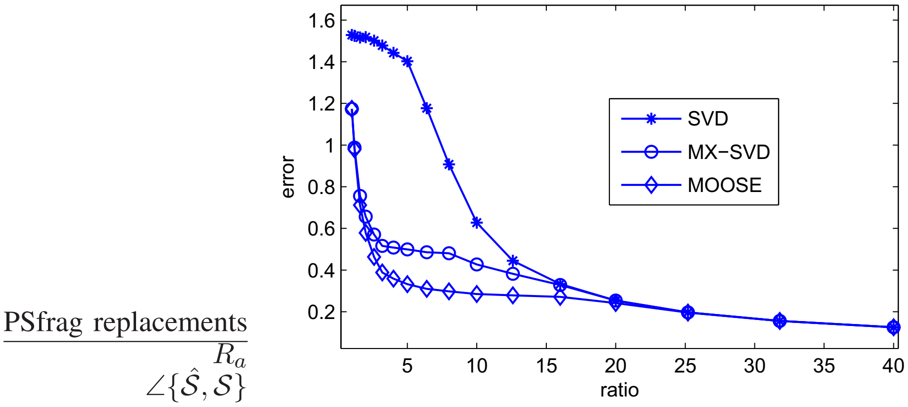

Fig. 4. **Mean subspace error vs. anomaly loading ratio R_a for parameters of Table I.** Mean-sample of the subspace error as a function of R_a obtained via a Monte-Carlo simulation using SVD (line with star marks), MX-SVD (line with circle marks), and MOOSE approach (line with diamond marks).

V. Real data simulation results

In this section we compare the performance of SVD, MX-SVD, MOOSE and HySime when applied to 4 hyperspectral image cubes. The images were collected by an AISA airborne sensor [25] configured to 65 spectral bands, uniformly covering VNIR range of $400nm$ - $1000nm$ wavelengths. The obtained image cubes are $b \times r \times c = 65 \times 300 \times 479$ hyperspectral images, where b, r and c denote the number of hyperspectral bands, the number of rows and the number of columns in the image cube, respectively.

The assumed signal-subspace rank is $k = 10$. It was deliberately chosen to be below the real signal subspace rank, the estimated values of which were found to be between 15-20, as obtained by applying MOCA on the images under evaluation. This poses signal-subspace estimation algorithms in challenging conditions, since by using a lower rank, we make the background vectors and the rare vectors compete harder for a better representation by the estimated signal subspace. This situation may occur in practical situations (such as local anomaly detection algorithms) where, on one hand, the application is optimized to work better in a low dimensional subspace, while on the other hand, this subspace is required to contain anomaly-related information.

The only ground-truth information available for this evaluation were locations of man-made objects. In Fig. 5 are shown images of the 30th-band of each of the 4 image cubes used for the evaluation. The ground-truth anomalies, which are marked in white and encircled by red ellipses, were manually identified using side information collected from high resolution RGB images of the corresponding scenes. The ground truth anomalies consist of vehicles and small agriculture facilities, which occupy few-pixel segments.

Since the man-made objects are anomalous in these images, it is difficult to represent them with low error by employing the classical ℓ_2-norm based methods, we evaluate the anomaly-preserving algorithm performances in terms of the maximum residual norms obtained on the ground-truth anomalies. That is, the best algorithm should have the following property: once applied on a whole image cube, the $\ell_{2,\infty}$-norm of the ground-truth anomaly residuals and the $\ell_{2,\infty}$-norm of the whole image should be the lowest compared to the other algorithm results obtained in all image cubes. In other words, the better algorithm represents better not only all image pixels, but also the anomalous ones.

Thus, in Table II one can see that MOOSE has the lowest $\ell_{2,\infty}$-norm of image residuals and the lowest $\ell_{2,\infty}$-norm of the ground-truth anomalies in all examined images. SVD and HySime have the highest $\ell_{2,\infty}$-norms of image residuals and anomaly residuals that are equal in all images, with a little advantage of HySime for most of the images. This shows that ℓ_2-based approaches poorly represent anomalies and that the worst-case error obtained by SVD and HySime in the whole image is on anomalies. The $\ell_{2,\infty}$-norms of image residuals and anomaly residuals obtained by MOOSE are different, meaning that the $\ell_{2,\infty}$-norms of image residuals are obtained on the background, i.e., the anomalies were represented even better than the background. It is instructive to note that the total CPU time consumed by MOOSE in our evaluations was twice as long as the CPU time consumed by MX-SVD (which is used by MOOSE for initialization). Since the results of MX-SVD are much better than those of SVD and comparable to those of MOOSE, it turns out that practically, MX-SVD is a good choice when one is looking for an anomaly preserving subspace estimator.

TABLE II

SUBSPACE ESTIMATION METHODS IN TERMS OF MAX. ERROR NORM

	Global $\ell_{2,\infty}$-norm of residuals				Anomaly $\ell_{2,\infty}$-norm of residuals			
Cube	SVD	HySime	MX-SVD	MOOSE	SVD	HySime	MX-SVD	MOOSE
1	200.6	180.6	98.3	97.3	200.6	180.6	82.7	81.7
2	1880.8	1854.9	312.5	282.0	1880.8	1854.9	312.5	207.8
3	453.0	403.0	98.5	73.6	453.0	403.0	84.1	70.9
4	749.6	755.1	445.6	401.2	749.6	755.1	445.6	383.6

VI. CONCLUSION

In this work we have proposed an algorithm for dimensionality reduction of high-dimensional noisy data that preserves rare-vectors. The proposed algorithm is optimal in the sense that the estimated subspace (locally) minimizes the maximal-norm of misrepresentation residuals. The optimization is performed via a natural conjugate gradient learning approach carried out on the set of n dimensional subspaces in $\mathbb{R}^m$, $m > n$, known as the Grassmann manifold. The proposed algorithm is denoted as *Maximum of Orthogonal complements Optimal Subspace Estimation* (MOOSE) and is the optimal version of a recently proposed greedy algorithm named *Min-Max-SVD* (MX-SVD). As any local minimization of a non-convex objective function, MOOSE is prone to getting trapped in a local minimum. Therefore, a proper initialization is crucial and is obtained by employing MX-SVD that uses global principles to find a suboptimal solution that is close to the global minimum. The results of MOOSE and MX-SVD were compared to the results of ℓ_2 - based techniques (SVD and HySime) by applying them both on simulated data and on real hyperspectral images. It was demonstrated that the results of MOOSE and MX-SVD are much better than those of SVD in terms of max-norm residual error, obtained in both simulated and real data, and in terms of the subspace estimation error obtained for simulated data. Although MX-SVD exhibits results inferior to those of MOOSE, the results of MX-SVD are quite comparable to those of MOOSE meaning that practically, the greedy MX-SVD algorithm is a good choice, since it is more computationally efficient.

ACKNOWLEDGMENT

The authors would like to thank Dr. Jose M.P. Nascimento for providing his HySime code that helped us to perform the comparative analysis of the proposed algorithm.

REFERENCES

[1] G.F. Hughes "On The Mean Accuracy Of Statistical Pattem Recognizers," *IEEE Trans. Infor. Theory, vol. IT-14, NO. I , pp 55 – 63, 1968.*

[2] M. Tipping and C. Bishop "Probabilistic principal component analysis", *Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol.61, Number 3, 1999, pp. 611-622.*

[3] O. Kuybeda, D. Malah and M. Barzohar "Rank Estimation and Redundancy Reduction of High-Dimensional Noisy Signals with Preservation of Rare Vectors," *IEEE Trans. Signal Proc., vol. 55, no. 12, pp. 5579-5592, Dec. 2007.*

[4] O. Kuybeda, D. Malah and M. Barzohar, Global Unsupervised Anomaly Extraction and Discrimination in Hyperspectral Images via Maximum- Orthogonal Complement Analysis, *EUSIPCO - European Signal Processing Conference, Aug. 2008, Lausanne. Switzerland.*

[5] A. Edelman, T. A. Arias and S. T. Smith "The Geometry of Algorithms with Orthogonality Constraints," *Siam J. Matrix Anal. Appl., vol. 20, no. 2, pp. 303-353, 1998.*

[6] A. Edelman, T. A. Arias and S. T. Smith "A Comprehensive Introduction to Differential Geometry," *vols. 13, 2nd ed., Publish or Perish, Houston, TX, 1979.*

[7] J. A. Snyman "Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms," *Springer Publishing (2005).*

[8] H. G. Grassmann “Die Ausdehnungslehre,” *Enslin, Berlin, 1862.*

[9] G. H. Golub and D. P. OLeary “Some History of the Conjugate Gradient and Lanczos Algo- rithms,” *19481976, SIAM Review 31 (1989), no. 1, pp. 50102.*

[10] A. Björck and G. H. Golub “Numerical Methods for Computing Angles Between Linear Subspaces,” *Mathematics of Computataion, vol. 27, no. 123, July 1973*

[11] G.W. Stewart *Matrix Algorithms Volume II: Eigensystems,* SIAM, Philadelphia, PA, 2001.

[12] S. T. Smith “Optimization techniques on Riemannian manifolds,” *Fields Institute Communications, vol. 3, AMS, Providence, RI, 1994, pp. 113146*

[13] L. L. Scharf *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis,* Addison-Welsey Publishing Company, 1993.

[14] G. W. Stewart and J. G. Sun, Matrix Perturbation Theory, Academic Press, Boston, MA, 1990.

[15] Jose M.P. Nascimento, Jose M.B. Dias, “Signal Subspace Identification in Hyperspectral Linear Mixtures”, *Lecture Notes in Computer Science, Vol.3523, Jan 2005, pp. 207 - 214*

[16] Q. Du, C. I. Chang, “A signal-decomposed and interference-annihilated approach to hyperspectral target detection” *Geoscience and Remote Sensing, IEEE Transactions on Vol.42, Issue 4, April 2004 pp. 892 - 906.*

[17] P. V. Overshee and B. D. Moor, “Subspace algorithms for the stochastic identification problem” *Automatica, vol. 29, pp. 649 - 660, 1993.*

[18] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue “Subspace methods for the blind identification of multichannel FIR filters,” *IEEE Trans. Signal Processing, vol. 43, pp. 516 - 526, Feb. 1995.*

[19] M. Viberg, “Subspace-based methods for the identification of linear time-invariant systems,” *Automatica, vol. 31, no. 12, pp. 1835 - 1853, 1995.*

[20] G. Xu, H. Liu, L. Tong, and T. Kailath, “A least-squares approach to blind channel identification” *IEEE Trans. Signal Processing, vol. 43, pp. 2982 2993, Dec. 1995.*

[21] M. E. Winter and E. M. Winter, “Comparison of approaches for determinig end-members in hyperspectral data,” *Aerospace Conference Proceedings, IEEE*, vol. 3, pp. 305 - 313, March 2000.

[22] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics., 2001.

[23] S. Boyd, L. Vandenberghe Convex Optimization Cambridge University Press, March 2004.

[24] C. T. Kelley, Iterative Methods for Optimization, Siam, 1999.

[25] Specim, Spectral Imaging LTD., *www.specim.fi*

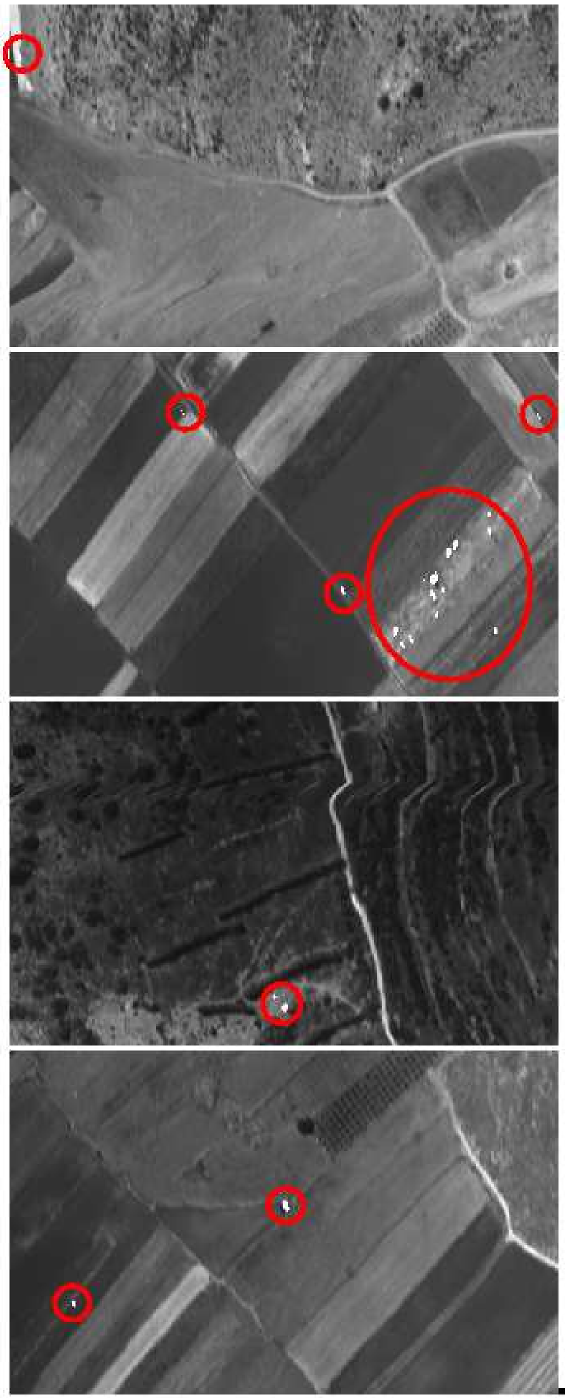

Fig. 5. **Ground truth.** A 30th-band of each one of 4 image cubes used for evaluation. The ground-truth anomalies were manually identified, marked in white and encircled in red.