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

Oleg Kuybeda, David Malah, Life Fellow, IEEE, 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 $\ell_2$-norm of the signal as compared to the noise, $\ell_2$-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 $\ell_2$-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 minimization 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 $\ell_2$- base approaches are demonstrated both on simulated and real hyperspectral data.

Index Terms—Anomaly detection, dimensionality reduction, Grassmann manifold, hyperspectral images, hyperspectral signal identification by minimum error (HySime), maximum orthogonal-complements analysis (MOCA), Min-Max-SVD (MX-SVD), redundancy reduction, signal-subspace rank, singular value decomposition (SVD).

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

$$x_i = As_i + z_i, \quad i = 1, \ldots, N$$  \hspace{1cm} (1)

where $x_i \in \mathbb{R}^p$ is the observed vector, $z_i \in \mathbb{R}^p$ is the data-acquisition or/and model noise; $s_i \in \mathbb{R}^r$, and $A \in \mathbb{R}^{p \times r}$ 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 $A$ are the pure materials spectra (end members) and $s_i$ their corresponding abundances [21],

A number of approaches have been proposed in the literature (e.g., [17]–[19]) for signal-subspace estimation under the assumption that $s_i$ and $z_i$ are independent, stationary, zero-mean and Gaussian. It was shown in [2] that for white noise $z_i$, 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 $\ell_2$-norm of misrepresentation residuals belonging to the complementary subspace, which can be obtained via singular value decomposition (SVD) of $X$, which is a matrix of observed data vectors $\{x_i\}$ ordered as its columns. The authors of [15], propose a new $\ell_2$-based approach, named as HySime, designed to determine both the signal subspace and

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

Manuscript received August 21, 2008; accepted August 24, 2009. First published September 22, 2009; current version published January 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tryphon T. Georgiou.

The authors are with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel (e-mail: kuybeda@gmail.com; malah@ee.technion.ac.il; meirb@vision-sense.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2032580

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on May 05,2010 at 14:45:24 UTC from IEEE Xplore. Restrictions apply.
its rank in hyperspectral imagery. The method first estimates the signal and noise covariance matrices. Then, they use the assumption on the nonnegativity of \( \{x_i\} \) in order to estimate the signal subspace rank by finding the subset of eigenvalues that best represents, in the \( \ell_2 \)-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 \( \ell_2 \)-based criterion is unsatisfactory for obtaining a reliable representation of the anomaly (rare) vectors, which typically contribute weakly to the \( \ell_2 \)-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 end members that are present in just a few data pixels contribute weakly to the \( \ell_2 \)-norm of the signal. Therefore, their contribution to the signal subspace cannot be reliably estimated using an \( \ell_2 \)-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 frequence in the data. Mathematically, the \( \ell_{2,\infty} \)-norm of a matrix \( X \) is defined as follows:

\[
\|X\|_{2,\infty} \triangleq \max_{i=1,\ldots,d} \|x_i\|_2
\]

where \( x_i \) denote columns of \( X \). In words: \( \|X\|_{2,\infty} \) means the maximum of \( \ell_2 \)-norms of \( X \) 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 the \( \ell_{2,\infty} \)-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 \( \mathbb{R}^m \), \( n \leq m \). Since the \( \ell_{2,\infty} \)-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 data (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 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 \( \tilde{S}_k \) should satisfy

\[
\tilde{S}_k = \arg\min_{\mathcal{L}} \|P_{\mathcal{L}^\perp} X\|_{2,\infty}^2
\]

s.t. rank \( \mathcal{L} = k \)

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

\[
\tilde{S}_k = \text{range} \left[ \Psi_{k-h} \Omega_h \right]
\]

where \( \Omega_h \) is a matrix composed of \( h \) columns selected from \( \chi \), and \( \Psi_{k-h} \) is a matrix with \( k-h \) orthogonal columns, obtained via SVD of \( P_{\Omega_h} X \). The main idea of MX-SVD is to collect anomaly vectors into \( \Omega_h \) in order to directly represent the anomaly vectors subspace. Since anomaly vectors are not necessarily orthogonal to background vectors, the matrix \( \Omega_h \) partially represents background vectors. The residual background vector contribution to the null-space of \( \Omega_h^T \) is represented by principal vectors found by applying SVD on \( P_{\Omega_h} X \).

The determination of the basis vectors of \( \tilde{S}_k \), in terms of \( \left\{ \Psi_{k-h} \Omega_h \right\} \) is performed as follows: First, we initialize \( \left[ \Psi_k \right] \Omega_0 \), such that

\[
\Psi_k = [u_1, \ldots, u_k] ; \quad \Omega_0 = []
\]

where \( u_1, \ldots, u_k \) are \( k \) principal left singular vectors of \( X \). Then, a series of matrices \( \left[ \left[ \Psi_{k-j} \right] \Omega_j \right] \) for \( j = 0 \) is constructed such that

\[
\Omega_{j+1} = \left[ \Omega_j \ x_{\omega_j} \right]
\]

\[
\Psi_{k-j-1} = [\psi_1, \ldots, \psi_{k-j-1}]
\]

where, for each \( i = 0, \ldots, k-1 \), \( \omega_i \) is the index of a data vector \( x_{\omega_j} \) that has the maximal residual squared norm \( r_i \):

\[
\omega_i = \arg\max_{n=1,\ldots,N} \| P_{\Omega_{j-i}} x_n \|
\]

\[
r_i = \| P_{\Omega_{k-i}} x_{\omega_j} \|^2
\]

and \( \psi_1, \ldots, \psi_{k-j-1} \) are \( k-i-1 \) principal left singular vectors of \( P_{\Omega_j} X \). Thus, the \( k \) columns of \( \left[ \Psi_{k-j} \right] \Omega_j \), 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 \( \Omega_j \) and by SVD-based vectors that best represent (in \( \ell_2 \) sense) the data residuals in the null-space of \( \Omega_j^T \). Moreover, each subspace is characterized by its maximum-norm misrepresentation residual \( r_j \). The greedy signal-subspace estimation
and $[W]$ is an equivalence class of all $p \times (p-k)$ orthogonal matrices whose columns span the same subspace in $\mathbb{R}^p$ as $W$. Here $[W]$ represents the orthogonal complement subspace to the sought signal subspace $S_k$. The set of all $n$-dimensional subspaces in $\mathbb{R}^m$, 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 $[W]$, say $W$, in order to be able to work with $[W]$ on the computer. Thus, by smoothly changing $W$, such that $[W] \in G_{p,p-k}$ we would be able to continuously move from one subspace to another and iteratively improve the objective function in a manner 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 $F([W])$ of (12).

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

$$\langle F_{[W]}; T \rangle = \langle \nabla F; T \rangle$$  

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

$$\langle \Delta_1, \Delta_2 \rangle = \text{tr}(\Delta_1^T \Delta_2).$$

(14)

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

$$T = W_\perp B$$

(15)

where $B$ are arbitrary $k \times k$ matrices and $W_\perp$ is a $p \times k$ orthogonal matrix satisfying

$$WW^T + W_\perp W_\perp^T = I$$

(16)

one obtains

$$\nabla F = F_{[W]} - W W^T F_{[W]}.$$  

(17)

III. MINIMIZING $\ell_{2,\infty}$-NORM ON THE GRASSMANN MANIFOLD

A. Problem Formulation

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

$$\hat{S} \in \arg\min_{[W]} F([W])$$

(11)

where the objective function $F([W])$ is defined as

$$F([W]) \triangleq \|W^T X\|_{\ell_{2,\infty}}^2$$

(12)

Fig. 1. MX-SVD flowchart. For a given signal subspace rank value $k$, constructs a signal-subspace basis of the form $S_k = \{\Psi_{k-h}; \Omega_h\}$, $h \in \{0, 1, \ldots, k\}$, which approximately minimizes $\|P_{\Omega_h} X\|_{\ell_{\infty}}$, where $\Omega_h$ is responsible for representing anomaly-vectors and, partially, background vectors; $\Psi_{k-h}$ complements $\Omega_h$ to represent background vectors in the $\ell_2$-sense.

$\hat{S}_k$ is selected as in (4), with

$$h = \arg\min_{j=0, \ldots, k} r_j$$

(10)

This policy combines the $\ell_2$-based minimization of background vector-residual norms with $\ell_\infty$-based minimization of anomaly vector residual norms, which produces a greedy estimate $\hat{S}_k$ that approximately satisfies (3). A flowchart summarizing the MX-SVD process is shown in Fig. 1.
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

$$\mathbf{W}(t) = (\mathbf{WVU}) \left( \frac{\cos(t\Sigma)}{\sin(t\Sigma)} \right) \mathbf{V}^T \quad (18)$$

where $t$ is a geodesic curve traversing parameter and $\mathbf{U\Sigma V}^T$ 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 $\Sigma$. It can be easily verified that the diagonal elements of the matrix $\mathbf{tU}$ 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 $\Sigma$ 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]

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

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 $\Sigma$. 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 $\sqrt{\text{tr}(\Sigma^2)}$ and, therefore, may change from iteration to iteration. In order to make it constant during the line search, for all iterations, the matrix $\Sigma$ is normalized:

$$\Sigma \triangleq \frac{\Sigma}{\sqrt{\text{tr}(\Sigma^2)}}. \quad (20)$$

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 small 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 line-search 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 $\beta$ in the Backtracking “while” stage, until the following condition holds:

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

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:

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

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 $\beta t$ 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 $\nabla F$ 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.

1) 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:

$$F_{\mathbf{W}} = \mathbf{x_jx_j}^T \mathbf{W} \quad (24)$$

where $\mathbf{x_j}$ is the vector for which $\max_{i=1,\ldots,N} ||\mathbf{W}_i^T \mathbf{x_i}||^2$ is obtained.
2) Case 2: If the maximum is obtained for a set of indexes \( J \) that contains more than one index, then the gradient direction \( \nabla \mathbf{G} = \mathbf{F}_W / \| \mathbf{F}_W \|_2 \) is given by solving the following problem:

\[
\nabla \mathbf{G} = \max_{\mathbf{G}} \min_{j \in J} \langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^T \mathbf{W} \rangle
\]

\[
\text{s.t.} \quad \langle \mathbf{G}, \mathbf{G} \rangle = 1
\]

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

\[
\| \mathbf{F}_W \| = \min_{j \in J} \langle \mathbf{G}, \mathbf{x}_j \mathbf{x}_j^T \mathbf{W} \rangle
\]

i.e., it equals to the lowest descent rate of the maximal residual norms, or equivalently, to the descent rate of } \| \mathbf{F}(\| \mathbf{W} \|) \text{ in the direction } \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 } \| \mathbf{F}(\| \mathbf{W} \|) \text{ along the geodesic given by } -\nabla \mathbf{F}, \text{ 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} \|), \text{ 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}

\[
\mathbf{H}_{s+1} = -\nabla \mathbf{F}_{s+1} + \gamma_s \hat{\mathbf{H}}_s
\]

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

\[
\gamma_s = \frac{\langle \nabla \mathbf{F}_{s+1} - \nabla \mathbf{F}_s, \nabla \mathbf{F}_{s+1} \rangle}{\langle \nabla \mathbf{F}_s, \nabla \mathbf{F}_s \rangle}
\]

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

\[
\hat{\mathbf{H}}_s = -\mathbf{W}_s \mathbf{V} \sin(\mathbf{\hat{\Sigma}}) + \mathbf{U} \cos(\mathbf{\hat{\Sigma}}) \mathbf{\Sigma}^T
\]

\[
\hat{\nabla} \mathbf{F}_s = \nabla \mathbf{F}_s - (\mathbf{W}_s \mathbf{V} \sin(\mathbf{\hat{\Sigma}}) + \mathbf{U} (1 - \cos(\mathbf{\hat{\Sigma}})) \mathbf{U}^T) \nabla \mathbf{F}_s.
\]

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, \text{ one has to perform a sequence of } d \text{ steps to get to a distance within } O(c^2) \text{ from the solution. However, in our problem it is not guaranteed that the } \ell_{2,\infty} \text{-based cost function is locally quadratic. Therefore, there is no guarantee that the conjugate gradient descent procedure converges in } d \text{ 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 nonconvex 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 } \| \mathbf{F}(\| \mathbf{W} \|) \text{ is summarized in Algorithm 2.
Algorithm 2: Conjugate Gradient Algorithm for Minimizing F([W]) on the Grassmann Manifold.

1) Given W_0, such that W_0^T W_0 = I_{p-k} and column space that coincides with the subspace obtained by MX-SVD, compute

F_{W_0} = x_j x_j^T W_0, with j satisfying ||W_0^T x_j||^2 = ||W_0^T X||^2_{2,\infty}
\n∇F_0 = F_{W_0} - W_0 W_0^T F_{W_0} and set H_0 = -∇F_0

2) For s = 0, 1, ..., 

2.1) Obtain the compact decomposition of H_s,
H_s = UΣV^T
2.2) Normalize the principal angles \tilde{Σ} = Σ/\sqrt{tr(Σ^2)}
2.3) Perform Backtracking-Armijo line search (see Algorithm 1) along the geodesic
W(t) = W_s Vcos(\tilde{Σ})V^T + Usin(\tilde{Σ})V^T

2.4) Update the subspace W_{s+1} = W(t)
2.5) Parallel transport the tangent vectors \dot{H}_s and \dot{∇}_s to the point [W_{s+1}]
\dot{H}_s = (\dot{W}_s Vsin(\tilde{Σ}) + Ucos(\tilde{Σ})) ΣV^T
\dot{∇}_s = ∇_s - (W_s Vsin(\tilde{Σ}) + Ucos(\tilde{Σ})) ΣV^T + U(I - cos(\tilde{Σ})) U^T ∇_s

2.6) Compute the new gradients
Euclidean: \dot{F}_{W_{s+1}} = x_j x_j^T W_{s+1}, with j satisfying ||W_{s+1}^T X||^2 = ||W_{s+1}^T X||^2_{2,\infty}
Grassmann: ∇F_{s+1} = \dot{F}_{W_{s+1}} + W_{s+1} W_{s+1}^T F_{W_{s+1}}

2.7) Compute the new search direction via Polak Ribiére conjugacy condition formula
H_{s+1} = -∇F_{s+1} + γ_s \tilde{H}_s, where
γ_s = (\dot{∇F}_{s+1} - \dot{∇F}_s, ∇F_{s+1}) / (∇F_s, ∇F_s)

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:

X = Y + Z

\begin{equation}
Y = \begin{bmatrix}
\sqrt{SNR_b} B S_b & \sqrt{SNR_a} A S_a
\end{bmatrix}
\end{equation}

where B is a p x r_b matrix with orthogonal unit-norm columns spanning the background subspace; A is a p x r_a matrix with orthogonal unit-norm columns spanning the subspace of anomalies; S_b is a r_b x N_b matrix of background vector coefficients with columns drawn randomly from a Gaussian distribution with covariance matrix C_b = I/r_b; S_a is a r_a x N_a matrix of anomaly vector coefficients with columns drawn randomly from a Gaussian distribution and normalized to have unit-norm; and Z is a p x (N_a + N_b) matrix containing white Gaussian noise with variance equal to 1/p.

| Table I: Maximum Residual-Norm Simulation Parameters |
|---|---|---|---|---|---|---|
| p | r_b | r_a | N_b | N_a | SNR_b | SNR_a |
| 100 | 5 | 5 | 10^3 | 10 | 100 | 10 |

For SNR defined as

SNR = \frac{E(||Y||^2)}{E(||Z||^2)}, \quad (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 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 ≪ N_b. However, their number is allowed to be higher than the anomaly subspace dimension that they belong to, i.e., N_a ≥ r_a. The extent of anomaly subspace population (loading) can be characterized by the loading ratio defined as follows:

R_a = \frac{N_a}{r_a}. \quad (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.

In Fig. 3 one can see empirical pdfs of the maximum-residual norm ||W^T X||^2_{2,\infty} obtained via a Monte Carlo simulation, where 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

||W^T X||^2_{2,\infty} \approx ||W^T Z||^2_{2,\infty}. \quad (34)

In words, the empirical distribution of the maximum data residual norm ||W^T X||^2_{2,\infty} for R_a = 1 is very close to the distribution of the maximum residual norm of noise ||W^T 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 because MX-SVD estimates the anomaly.
subspace by directly selecting 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 dotted–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.

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 \( \angle \{ \hat{S}, S \} \) defined as follows [11]:

\[
\angle \{ \hat{S}, S \} = \max_{\mathbf{u} \in \hat{S}} \min_{\mathbf{v} \in S} \angle \{ \mathbf{u}, \mathbf{v} \}, \quad \mathbf{u} \neq 0, \mathbf{v} \neq 0
\]

where \( \hat{S} \) and \( S \) 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 \([1, 4]\) 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 MOOSE approach (line with diamond marks). As clearly seen from the figure, the proposed approach corresponds to the lowest mean

---

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

**Fig. 4.** Mean subspace error versus 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).

**Fig. 5.** Ground truth. A thirtieth-band of each one of four image cubes used for evaluation. The ground-truth anomalies were manually identified, marked in white and encircled in red.
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 $\ell_2$-norm.

### 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 400–1000-nm wavelengths. The obtained image cubes are $b \times r \times c = 65 \times 300 \times 470$ 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 and 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 thirtieth-band of each of the four 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 $\ell_2$-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 $\ell_2$-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.

### 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 nonconvex 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 $\ell_2$ – 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

<table>
<thead>
<tr>
<th>Cube</th>
<th>SVD</th>
<th>HySime</th>
<th>MX-SVD</th>
<th>MOOSE</th>
<th>SVD</th>
<th>HySime</th>
<th>MX-SVD</th>
<th>MOOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200.6</td>
<td>180.6</td>
<td>98.3</td>
<td>97.3</td>
<td>200.6</td>
<td>180.6</td>
<td>82.7</td>
<td>81.7</td>
</tr>
<tr>
<td>2</td>
<td>1880.8</td>
<td>1854.9</td>
<td>312.5</td>
<td>282.0</td>
<td>1880.8</td>
<td>1854.9</td>
<td>312.5</td>
<td>207.8</td>
</tr>
<tr>
<td>3</td>
<td>453.0</td>
<td>403.0</td>
<td>98.5</td>
<td>73.6</td>
<td>453.0</td>
<td>403.0</td>
<td>84.1</td>
<td>70.9</td>
</tr>
<tr>
<td>4</td>
<td>749.6</td>
<td>755.1</td>
<td>445.6</td>
<td>401.2</td>
<td>749.6</td>
<td>755.1</td>
<td>445.6</td>
<td>383.6</td>
</tr>
</tbody>
</table>
MOOSE, the results of MX-SVD are quite comparable to those of MOOSE meaning that practically, the greedy MX-SVD MOOSE, the results of MX-SVD are quite comparable to those providing his HySime code that helped us to perform the comparative analysis of the proposed algorithm.

ACKNOWLEDGMENT

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

REFERENCES


David Malah (S’67–M’71–SM’84–F’87–LF’09) received the B.Sc. and M.Sc. degrees from the Technion—Israel Institute of Technology, Haifa, Israel, in 1964 and 1967, respectively, and the Ph.D. degree from the University of Minnesota, Minneapolis, in 1971, all in electrical engineering.

Following one year on the staff of the Electrical Engineering Department of the University of New Brunswick, Fredericton, NB, Canada, he joined the Technion in 1972, where he is an Elron-Elbit Professor of Electrical Engineering. From 1979 to 2001, he spent approximately six years, cumulatively, of sabbaticals and summer leaves at AT&T Bell Laboratories, Murray Hill, NJ, and AT&T Labs, Florham Park, NJ, conducting research in the areas of Speech and Image Communication, and summer 2004 at the Georgia Centers for Advanced Telecommunications Technology—GCAT, working in the area of video processing. Since 1975, he has been the academic head of the Signal and Image Processing Laboratory (SIPL), at the Technion, which is active in Image/Video and Speech/Audio Processing research and education. Since 2006, he has been the Director of the Center for Communication and Information Technologies—CCIT, at the Electrical Engineering Department, the Technion. His main research interests are in image, video, speech, and audio coding; speech and image enhancement; hyperspectral image analysis; data embedding in signals, speech synthesis and voice conversion, and in digital signal processing techniques.

Meir Barzohar received the B.S. and M.S. degrees in electrical engineering from the Technion—Israel Institute of Technology in 1973 and 1978, respectively, and the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from Brown University, Providence, RI, in 1993, working on new geometric, algebraic, and probabilistic models for finding and recognizing roads and their features in aerial images.

He has been employed by Rafael Israel since 1974 as a Research Engineer, in the areas of communication, digital signal processing, image processing, and image compression. In 1982, he was on a sabbatical at RCA Laboratories, Princeton, NJ, working on digital video signal processing for FM transmission and television scrambler for which he received a patent. From 1989 to 1994, he was granted a sabbatical and leave of absence from Rafael to join Brown University to pursue the Ph.D. degree and later a postdoctoral position. From 1994 to 2001, he was a Senior Research Scientist in the computer vision group in Rafael, working on probabilistic models for detection and tracking objects in image sequences. Since 2001, he has been at Visionsense medical company, Petah Tikva, Israel; as a group leader in the image processing group, working on 3-D geometry camera models, color image processing for development of stereoscopic endoscope, and 3-D reconstruction from stereo images, and 3-D-to–2-D registration for fusion of different modalities in the medical field. He is also a co-adviser of M.Sc. and Ph.D. students at the Electrical Engineering Department of the Technion. His research interests are in computer vision, 2-D and 3-D object recognition, Bayesian estimation and decision theoretic framework for image segmentation, hyperspectral image analysis and representation approaches, image compression, and image processing.