CCIT Report #623
May 2007

Rank Estimation and Redundancy Reduction of High-Dimensional Noisy Signals with Preservation of Rare Vectors

Oleg Kuybeda, David Malah, and Meir Barzohar*

Department of Electrical Engineering

Technion IIT, Haifa 32000, Israel

koleg@techunix.technion.ac.il, malah@ee.technion.ac.il,

meirb@visionsense.com

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. For example, when applying redundancy reduction techniques to hyperspectral images, it is essential to preserve anomaly pixels for target detection purposes. Since rare-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. The proposed approach combines ℓ_2 and ℓ_∞ norms for both signal-subspace and rank determination and considers two aspects: One aspect deals with signal-subspace estimation aiming to minimize the *maximum* of data-residual ℓ_2-norms, denoted as $\ell_{2,\infty}$, for a given rank conjecture. The other determines whether the rank conjecture is valid for the obtained signal-subspace by applying Extreme Value Theory results to model the distribution of the noise $\ell_{2,\infty}$-norm. These two operations are performed alternately using a suboptimal greedy algorithm, which makes the proposed approach practically plausible. The algorithm was applied on both synthetically simulated data and on a real hyperspectral image producing better results than common ℓ_2-based methods.

Index Terms

Signal-subspace rank, singular value decomposition (SVD), minimum description length (MDL), anomaly detection, dimensionality reduction, redundancy reduction, hyperspectral images.

I. INTRODUCTION

Redundancy reduction is one of the central problems faced when dealing with high-dimensional noisy signals. In many sensor-array applications, signal vectors belong to a lower-dimensional subspace than the observed data. This signal-subspace could be estimated and used for redundancy reduction by projecting the observed data vectors onto it. The estimated signal-subspace properties should adequately reflect needs of application that uses this low-dimensional subspace. In this paper, we focus on applications that analyze anomaly vectors, such as target detection in hyperspectral images. Therefore, the estimated signal-subspace should contain (preserve) such vectors. The knowledge of signal-subspace implies also a knowledge of the corresponding signal-subspace rank. In a number of applications in the literature the signal rank (order) is assumed to be known - such as the number of independent source signals in Blind Source Separation via Independent Components Analysis [21]; the order of the channel FIR model in blind single-input/multiple-output channel identification [15], [16], [17]; the signal-subspace rank in linear system identification algorithms [7], [8], [10]; the number of individual pure spectra (endmembers) in hyperspectral image processing [24], etc.

In practice, the signal-subspace and rank have to be estimated from observed vectors $\{\mathbf{x}_i\}_{i=1}^N$, assumed to satisfy the following linear model:

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

where $\mathbf{x}_i \in \mathbb{R}^p$ is the observed random vector, $\mathbf{z}_i \in \mathbb{R}^p$ is the data-acquisition or/and model noise; $\mathbf{s}_i \in \mathbb{R}^r$, and $\mathbf{A} \in \mathbb{R}^{p\times r}$, ($r \leq p$). In some of the applications above, $\mathbf{s}_i$ is a vector of hidden source signals, $\mathbf{A}$ is some "mixing" matrix through which the sources are observed; while in a hyperspectral application the columns of $\mathbf{A}$ are the pure materials spectra (endmembers) and $\mathbf{s}_i$ their corresponding abundances [24]. The observed dimension p is obviously known, whereas the signal-intrinsic dimension (rank) r is not.

A number of approaches have been proposed in the literature (see [18], [19], [20]) for signal-subspace and rank estimation under the assumption that $\mathbf{s}_i$ and $\mathbf{z}_i$ are independent, stationary, zero-mean and ergodic random Gaussian processes. Under these assumptions, the estimated signal subspace is determined by minimizing the ℓ_2-norm of misrepresentation residuals belonging to the complementary subspace.The classical methods are principal components analysis (PCA) for signal-subspace estimation and methods like minimum description length (MDL) [11] and others [14] for rank determination.

In this paper, we propose a redundancy reduction approach for high-dimensional noisy signals con-

taining anomaly (rare) vectors that, typically, contribute weakly to the ℓ_2-norm of the signal as compared to the noise. This makes ℓ_2-based criteria unsatisfactory for obtaining a good representation of rare vectors, which 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 (anomalous) endmembers that are present in just a few data pixels contribute weakly to the ℓ_2-norm of the signal, compared to the noise. Therefore, their contribution to the signal-subspace cannot be reliably estimated using an ℓ_2-based criterion, as will be shown in more detail in the following sections. Yet, the representation of the rare vectors can be crucial for anomaly detection that might follow the redundancy reduction stage.

The problem of representing well and compactly all signal vectors, including rare ones, in a low-dimensional subspace didn't attain much attention in the literature. The opposite is true: there are applications where the rare-vectors are treated as outliers that may skew the nominal signal-subspace estimation. The problem of dealing with outliers has been extensively studied in the literature. Related works ([9],[22],[23] and many others) propose robust parameter estimation techniques, which are designed to exclude the outlying measurements.

In contrast to robust parameter estimation techniques, the method proposed here is designed to represent well both abundant and rare measurements, irrespective of their frequentness in the data. In other words, a good representation of all measured vectors is equally important. For this purpose, we define a deterministic matrix $\mathbf{Y} \in \mathbb{R}^{p\times N}$ that consists of signal components only. Our goal is to find the column space and the rank of $\mathbf{Y}$, given an observed matrix $\mathbf{X} \in \mathbb{R}^{p\times N}$ with columns $\mathbf{x}_1, \ldots, \mathbf{x}_N$,

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z}, \tag{2}$$

where $rank\ \mathbf{Y} = r$ is unknown, $r < p, N$, and $\mathbf{Z} \in \mathbb{R}^{p\times N}$ is a noise matrix with i.i.d. zero-mean Gaussian entries.

Our approach combines two norms, ℓ_2 and ℓ_∞ for both signal-subspace and rank determination and considers two aspects: One aspect deals with the determination of the signal-subspace for a given rank conjecture. The other determines whether the rank conjecture is valid, given the obtained signal-subspace. The corresponding operations are performed alternately for an increasing sequence of tested subspace rank values, until the rank conjecture is affirmed. The signal-subspace estimation aims to minimize the *maximum* of misrepresentation-residual ℓ_2-norms denoted as $\ell_{2,\infty}$-norm. Mathematically, the $\ell_{2,\infty}$-norm

of a matrix $\mathbf{X}$ is defined as follows:

$$\|\mathbf{X}\|_{2,\infty} \triangleq \max_{i=1,\ldots,N} \|\mathbf{x}_i\|_2, \tag{3}$$

where $\mathbf{x}_i$ denote columns of $\mathbf{X}$. It is easy to see that $\ell_{2,\infty}$ is a norm on a vector space $\mathcal{V}$ of $p \times N$ matrices, since for any $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2 \in \mathcal{V}$ the following holds:

1. $\|\alpha\mathbf{X}\|_{2,\infty} = |\alpha| \|\mathbf{X}\|_{2,\infty}$,
2. $\|\mathbf{X}_1 + \mathbf{X}_2\|_{2,\infty} \leq \max\limits_i \ (\|\mathbf{x}_{1,i}\|_2 + \|\mathbf{x}_{2,i}\|_2) \leq \max\limits_i \ \|\mathbf{x}_{1,i}\|_2 + \max\limits_i \ \|\mathbf{x}_{2,i}\|_2 = \|\mathbf{X}_1\|_{2,\infty} + \|\mathbf{X}_2\|_{2,\infty}$,
3. $\|\mathbf{X}\|_{2,\infty} \geq 0$,
4. $\|\mathbf{X}\|_{2,\infty} = 0 \iff \mathbf{X} = 0$.

The signal-subspace rank is determined by applying Extreme Value Theory results [25] to model the distribution of the misrepresentation $\ell_{2,\infty}$-norm. Since $\ell_{2,\infty}$ penalizes individual data-vector misrepresentations, it helps to represent well not only abundant-vectors, but also rare-vectors. Since we use the maximum-orthogonal-complements (residuals) for the determination of both signal-subspace and rank, we call the proposed algorithm: *Maximum Orthogonal-Complements Algorithm (MOCA).*

This paper is organized as follows: Section II discusses an optimality criterion for signal-subspace determination for data that contains rare-vectors. In Section III we describe a signal-subspace determination approach, which is based on a combination of ℓ_2 and $\ell_{2,\infty}$ norms. We denote this approach as Min-Max SVD (MX-SVD). In Section IV we show simulation results that compare the performances of classical SVD and the new MX-SVD approaches for signal-subspace determination in presence of anomaly vectors. In Section V we describe the signal-subspace rank determination approach that preserves rare-vectors. In Section VI we present simulation results of comparing the performance of classical MDL with the proposed approach for signal-subspace rank determination. The comparison is performed on both synthetically simulated data and on a real hyperspectral image. Finally, in section VII we conclude this work.

II. Optimality criterion for subspace determination

Before getting into the development of an estimator of a subspace that may include rare vectors, we first characterize the presence of rare-vectors. For demonstrational purposes, we show in Fig. 1 a schematic plot of a subspace of abundant vectors and rare-vectors. The abundant vectors (marked by dots) lie close to a subspace spanned by the vector $\mathbf{v}_0$. As it is seen in the figure, the rare vectors (marked by circles and dashed arrows) $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ don't belong to the abundant vector subspace spanned by $\mathbf{v}_0$. Obviously, rare-vectors are characterized by their low number compared to the number of abundant vectors. Rare

vectors are supposed to lie far from the abundant vector subspace. They, however, are allowed to belong to a subspace of a dimension lower than their number. It is important to stress that unlike in the example (for $p = 2$), the observed dimensionality p (in the general case) is expected to exceed the dimension of the subspace spanned by abundant and rare vectors combined.

The example above can be generalized by the following property:

Rare vector presence property: The $p \times N$ matrix $\mathbf{Y}$ is said to contain rare-vectors if there exists a decomposition $\mathbf{Y} = [\mathbf{Y}_1|\mathbf{Y}_2]\mathbf{\Pi}$, where $\mathbf{\Pi}$ is some permutation matrix, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are $p \times N_1$ and $p \times N_2$ submatrices of $\mathbf{Y}$, such that $N_1 + N_2 = N$, $N_1 \gg N_2$, and $range\ \mathbf{Y}_1 \subset range\ \mathbf{Y}$.

This property states that the matrix $\mathbf{Y}$ is considered to contain rare vectors if the number of $\mathbf{Y}$ columns that are linearly independent of all the other $\mathbf{Y}$ columns is relatively small.

In order to develop a rare-vector preserving signal-subspace estimator, we should define an optimality criterion that is sensitive to the appearance of rare-vectors in the data. First, we consider an ℓ_2-based optimality criterion, since it appears in Singular Value Decomposition (SVD) - a well-known technique for the signal-subspace estimation [3]. Then, we show that ℓ_2-based criteria are not appropriate for estimating a signal-subspace that contains rare vectors, and propose combining ℓ_2 and ℓ_∞-based criteria as a remedy.

As noted above, at the signal-subspace determination stage, the rank is assumed to be known, say $rank\ \mathbf{Y} = k$.

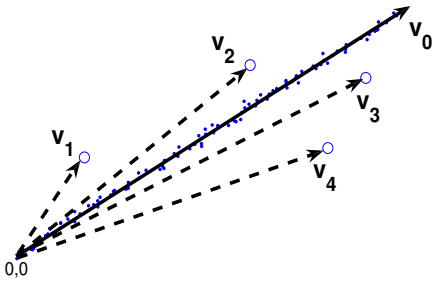


Fig. 1. **Schematic plot demonstrating rare vectors presence in data.** $\mathbf{v}_0$ spans abundant vectors (dots) subspace; $\mathbf{v}1, \mathbf{v}_2, \mathbf{v}_3$ and $\mathbf{v}_4$ denote rare vectors (circles).

A. Signal-subspace estimation via SVD

According to the SVD approach, the signal-subspace $\mathcal{S}_k = range\ \mathbf{Y}$ is estimated by minimizing the ℓ_2 norm of the residuals:

$$\begin{aligned} \hat{\mathcal{S}}_k &= \underset{\mathcal{L}}{\text{argmin}}\ \|\mathbf{X} - \mathcal{P}_{\mathcal{L}}\mathbf{X}\|_{Fb}^2 = \underset{\mathcal{L}}{\text{argmin}}\ \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{Fb}^2 \\ &\quad \text{s.t. } \textit{rank}\ \mathcal{L} = k, \end{aligned} \tag{4}$$

where $\|\cdot\|_F b$ denotes Frobenius norm, $\mathcal{L} \subset \mathbb{R}^p$ and $\mathcal{P}_{\mathcal{L}}$ denotes an orthogonal projection onto subspace $\mathcal{L}$. It can also be shown that under a Gaussian assumption on the columns of $\mathbf{Y}$, $\hat{\mathcal{S}}_k$ coincides with the maximum-likelihood (ML) estimation of $\mathcal{S}_k$ [1]. The estimated signal-subspace $\hat{\mathcal{S}}_k$ is obtained via SVD of the observation matrix $\mathbf{X}$ as $\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}'$, where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are $p \times p$ and $N \times p$ matrices with orthonormal columns, respectively, and $\hat{\mathbf{S}} =$ diag $\{\hat{s}_1, \ldots, \hat{s}_p\}$, $\hat{s}_1 \geq \hat{s}_2 \geq \ldots \geq \hat{s}_p$. The signal subspace $\hat{\mathcal{S}}_k$ is equal to span of $\{\hat{\mathbf{u}}_\mathbf{1}, \ldots, \hat{\mathbf{u}}_\mathbf{k}\}$ - the first k columns of $\hat{\mathbf{U}}$ (see [3] for details).

B. Drawbacks of minimizing the ℓ_2 norm in the presence of rare-vectors

Intuitively, it seems that minimizing the observation residuals $\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i$, $i = 1, \ldots N$, as a function of $\mathcal{L}$, could be appropriate for estimating $\mathcal{S}$. Indeed, for $\mathcal{L} = \mathcal{S}$,

$$\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i = \mathcal{P}_{\mathcal{L}^\perp}\mathbf{z}_i, \qquad i = 1, \ldots, N, \tag{5}$$

which means that given a precise signal-subspace estimation, the data residuals are equal to the corresponding noise residuals. Whereas, for $\mathcal{L} \neq \mathcal{S}$, one expects to obtain signal contributions in the residual subspace $\mathcal{L}^\perp$, which are likely to increase the residual squared norm $\|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i\|^2$. This can be seen from the fact that since $\mathbf{z}$ is statistically independent of $\mathbf{y}$, so are their projections onto the null-space of $\mathcal{L}$. Moreover, since z_i are zero mean i.i.d., it is expected that

$$\|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_i\|^2 \approx \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{y}_i\|^2 + \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{z}_i\|^2. \tag{6}$$

Therefore, looking for $\hat{\mathcal{S}}$ that minimizes residual norms is reasonable. However, using an ℓ_2 norm (like in (4)) can be inappropriate in the presence of rare-vectors, since the contribution of rare-vector residuals to the ℓ_2-norm may be much weaker than the contribution of noise-residuals. As a result, the estimated subspace $\hat{\mathcal{S}}$ may be skewed by noise in a way that completely misrepresents the rare-vectors. In some practical cases this miss-representation may occur with high probability, as demonstrated in simulations below.

First, we define the Rare-vector Signal-to-Noise Ratio as follows:

$$RSNR \triangleq \frac{s^2_{min}(\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{Y}_{rare})}{E\{\|\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{z}\|_2^2\}} = \frac{s^2_{min}(\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}\mathbf{Y}_{rare})}{(p-k)\sigma^2}, \tag{7}$$

where $\mathbf{Y}_{rare}$ is a submatrix of $\mathbf{Y}$ composed of all rare-vectors, $\mathbf{Y}_{abund}$ is a submatrix of $\mathbf{Y}$ composed of the remaining (abundant) vectors; $\mathcal{P}_{\mathbf{Y}^{\perp}_{abund}}$ is a projection onto the null-space of $\mathbf{Y}_{abund}$; $s^2_{min}(\mathbf{D})$ is the squared minimal non-zero singular value of the argument matrix $\mathbf{D}$, and σ^2 is the noise variance. The choice of the minimal non-zero singular value is essential, since it reflects the rare-vectors subspace perturbation by additive noise [4], i.e., the error in rare-vector subspace estimation. That is, RSNR measures the ratio between the contribution of rare-vectors in the direction of the least-significant eigenvector of the rare vector-residuals in the null-space of abundant vectors, and the contribution of noise in that direction. We also define SNR as follows:

$$SNR \triangleq \frac{\|\mathbf{Y}_{abund}\|^2_{Fb}}{pN\sigma^2}. \tag{8}$$

Now, we describe the setup of simulations that show a typical case for which rare-vectors are misrepresented by SVD. A $p \times N = 10^2 \times 10^5$ signal matrix $\mathbf{Y}$ (which corresponds to a typical hyperspectral image cube consisting of 10^5 pixel-vectors of dimension 10^2, each) was generated, such that $\mathbf{Y} = [\mathbf{Y}_{abund}|\mathbf{y}_{rare}]$, using a Gaussian distribution for the columns of $\{\mathbf{Y}_{abund}$ with a covariance matrix $\mathbf{C}_{\mathbf{Y}_{abund}} = 100\sigma^2\mathbf{I}_{p,q}$, $q = 5$, and $\mathbf{y}_{rare} \in null\ \mathbf{Y}^T_{abund}$, where $\mathbf{I}_{p,q}$ denotes a diagonal $p \times p$ matrix with $q \leq p$ nonzero diagonal entries, all equal to 1. Since $s^2_{min}(\mathbf{y}_{rare}) = \|\mathbf{y}_{rare}\|_2^2$, it follows that

$$RSNR = \frac{\|\mathbf{y}_{rare}\|_2^2}{E\{\|\mathcal{P}_{\mathbf{C}_{\mathbf{Y}_{abund}}{}^{\perp}}\mathbf{z}\|_2^2\}} = \frac{\|\mathbf{y}_{rare}\|_2^2}{(p-k)\sigma^2}, \tag{9}$$

Then, the "measured" data-matrix $\mathbf{X}$ was obtained as $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$, where the $\mathbf{Z}$ columns are Gaussian with a covariance matrix $\mathbf{C}_z = \sigma^2\mathbf{I}_{p,p}$.

In our simulations, for each $RSNR$ value $\mathbf{X}$ was generated 100 times. We consider 50 RSNR values, sampled uniformly in the range $[0, \ldots, 170]$ for $\sigma^2 = 1$, as shown in Fig. 2 (a). In a dashed (dot-dashed) line we plot the minimum (maximum) of 100 generated values (per RSNR value) of

$$\nu_k \triangleq \|\mathcal{P}_{\hat{\mathcal{S}}^{\perp}_k}\mathbf{X}\|^2_{2,\infty} \triangleq \max_{j=1,\ldots,N} \|\mathcal{P}_{\hat{\mathcal{S}}^{\perp}_k}\mathbf{x}_j\|_2^2, \tag{10}$$

where $\ell_{2,\infty}$ is a norm defined by selecting the *maximum ℓ_2-norm* of the data vector residuals (corresponding to the null-space of $\hat{\mathcal{S}}_k$, $k = q + 1 = 6$ in (10)). In a thin solid line we plot $\|\mathbf{y}_{rare}\|_2^2$ as a function of RSNR.

We repeated the simulation above for matrices $\mathbf{X}$ with $\mathbf{Y} = \mathbf{Y}_{rare}$ (i.e., there are no rare-vectors in the data). The horizontal heavy solid line shows the mean value of ν_k, $k = q = 5$, corresponding to data without rare-vectors. In both cases - with and without a rare-vector, $\hat{\mathcal{S}}_k$ was obtained via SVD.

The maximum residual norm $\nu_k = \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ in data without rare-vectors has a narrow distribution, since it approximately equals to the maximum norm of the noise residuals $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$, which has a narrow distribution, explained by Extreme Value Theory results, as shown in Appendix I. Therefore, ν_k has nearly a "deterministic" behavior in data without rare-vectors.

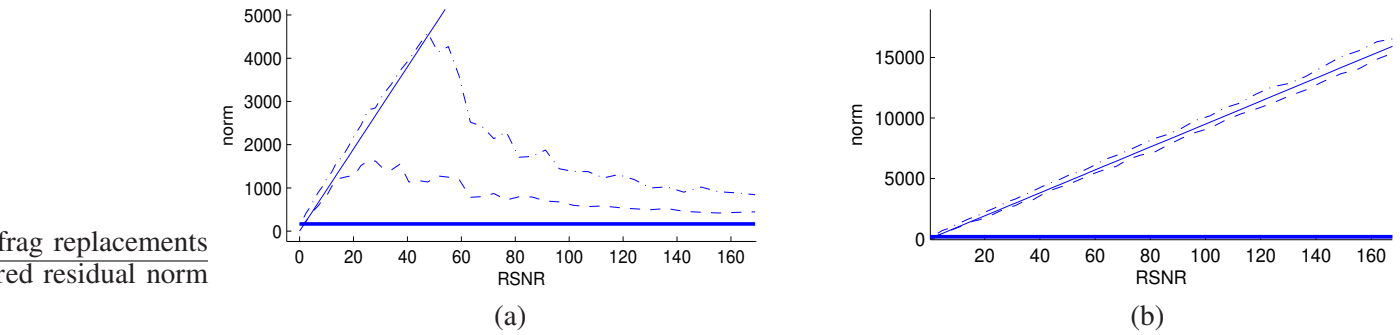
However, in the presence of rare-vectors (for $k = q+1$), it is likely to obtain ν_k values that are higher than $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$. Thus, as it is seen from the figure, there is a range of RSNR values ($0 < RSNR < 140$, $p = 10^2$ and $N = 10^5$ in this example), for which the value of ν_k in the presence of a rare-vector lies much higher (between the dot-dashed and dashed lines, representing the min and max values, respectively) than the nearly deterministic value of ν_k in the absence of a rare-vector (heavy horizontal solid line). This phenomenon corresponds to the poor representation of rare-vectors by SVD. This range of RSNR values, however, is of high practical importance in some applications. For instance, in hyperspectral that we examined, characterized by SNR $= 100$, the observed RSNR satisfies RSNR ≤ 30, which means that SVD would most likely fail to appropriately represent rare-vectors in this application. On the other hand, for high RSNR values, the rare-vector contributions becomes stronger in the ℓ_2-sense, compared to the noise contributions. As a result, for high RSNR values, SVD represents well the rare-vectors. This can be seen from the fact that the dot-dashed line in Fig. 2 converges to the heavy horizontal solid line.

For clarification, in Fig. 2 (b) we show results of the above simulation for an assumed incorrect dimensionality value of $k-1$. As expected, SVD "prefers" to represent abundant vectors. This results in a maximum misrepresentation error that is dictated solely by the norm of the rare-vector for a much wider range of $RSNR$ values. Note that the min and max values are not equal because of the noise added to the rare vector. We also simulated the case of a "wrong" dimensionality of $k+1$ and noticed, as expected, that it produces results close to the case of the correct dimensionality k in Fig. 2 (a).

In summary, the above example demonstrates that SVD may poorly represent the rare-vectors for an important range of low RSNR values.

C. Signal-Subspace determination by $\ell_{2,\infty}$-norm minimization

In the last example we have seen that SVD, being ℓ_2-optimal (4), may not be sensitive to rare-vectors, leaving large rare-vector residuals in $\hat{\mathcal{S}}_k^{\perp}$. In order to tackle this problem, we propose using $\ell_{2,\infty}$ instead of ℓ_2, which transforms the optimization problem (4) to the following form:



(a) (b)

Fig. 2. **Monte-Carlo simulation of SVD-based signal-subspace estimation in the presence of rare-vectors for $p = 10^2$ and $N = 10^5$.** The rare-vector squared norm $\|\mathbf{y}_{rare}\|_2^2$ (solid thin line), the sample-minimum of maximum data-residual squared-norms ν_k *in the presence* of rare-vectors (dashed line), the sample-maximum of the maximum data-residual squared norm ν_k *in the presence* of a rare-vector (dot-dashed line), the sample-mean of maximum noise-residual squared-norms ν_k *in the absence* of rare-vectors (heavy horizontal solid line); a) for correct rank k b) for "wrong" rank $k-1$.

$$\begin{aligned} \hat{\mathcal{S}}_k &= \underset{\mathcal{L}}{\text{argmin}} \ \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{X}\|_{2,\infty}^2 \\ &\quad \text{s.t.} \qquad \text{rank } \mathcal{L} = k. \end{aligned} \tag{11}$$

The objective function of this optimization problem is not differentiable and, therefore, is hard to optimize. In order to make the problem differentiable, analogously to the Chebyshev (minimax) approximation problem in [31], the problem of (11) can be recast as follows:

$$\begin{aligned} \hat{\mathcal{S}}_k &= \underset{\mathcal{L},\gamma}{\text{argmin}} \ \gamma \\ \text{s.t.} \qquad & \|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_j\|_2^2 \le \gamma \qquad \forall j = 1,\ldots,N, \\ & \text{rank } \mathcal{L} = k, \end{aligned} \tag{12}$$

where the additional parameter γ was introduced to bound all residual squared norms $\|\mathcal{P}_{\mathcal{L}^\perp}\mathbf{x}_j\|^2$ (including the maximal one) from above. Thus, by minimizing this bound with respect to $\mathcal{L}$, one minimizes the maximum residual norm corresponding to $\|\mathbf{X}\|_{2,\infty}$ of (11), which makes problems (11) and (12) equivalent.

Although the obtained equivalent optimization problem is differentiable, it still seems to be practically intractable because of the large multiplicity of constraints, which is equal to N (the number of data vectors). Therefore, in the next section we propose a suboptimal greedy algorithm that is found to produce good results.

III. Signal-Subspace determination by combining SVD with min-max of residual norms (MX-SVD)

In order to make the minimization of (11) or (12) computationally plausible, we propose to constrain the sought $\hat{\mathcal{S}}_k$ basis to be of the following form:

$$\hat{\mathcal{S}}_k = \text{range } [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h], \tag{13}$$

where $\mathbf{\Omega}_h$ is a matrix composed of h columns selected from $\mathbf{X}$, and $\mathbf{\Psi}_{k-h}$ is a matrix with $k-h$ orthogonal columns, obtained via SVD of $\mathcal{P}_{\mathbf{\Omega}_h^{\perp}}\mathbf{X}$. As demonstrated in the previous section, the $\ell_{2,\infty}$ norm of data-vector residuals is governed by the rare-vector miss-representations via SVD (which is ℓ_2 - optimal), whereas abundant vectors can be successfully represented via SVD. Therefore, the main idea of the proposed approach, which we denote as MX-SVD, is to collect rare-vectors into $\mathbf{\Omega}_h$ in order to directly represent the rare-vectors subspace. Since rare-vectors are not necessarily orthogonal to abundant vectors, matrix $\mathbf{\Omega}_h$ also partially represents *abundant vectors*. The residual abundant vector contribution to the null-space of $\mathbf{\Omega}_h^T$ is represented by principal vectors found by applying SVD on $\mathcal{P}_{\mathbf{\Omega}_h^{\perp}}\mathbf{X}$. As noted above, the columns in $\mathbf{\Omega}_h$ are directly selected from $\{\mathbf{x}_i\}_1^N$, the set of noisy data vectors. Although this makes $range\,\mathbf{\Omega}_h$ a noisy estimation of the pure rare-vectors subspace, it still represents well the noisy rare-vectors in the data, which is, actually, the main objective of MOCA.

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

$$\mathbf{\Psi}_k \;=\; [\mathbf{u}_1,\ldots,\mathbf{u}_k]\,;\; \mathbf{\Omega}_0 = [], \tag{14}$$

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

$$\mathbf{\Omega}_{i+1} \;=\; [\mathbf{\Omega}_i|\mathbf{x}_{\omega_i}] \tag{15}$$

$$\mathbf{\Psi}_{k-i-1} \;=\; \left[\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_{k-i-1}\right], \tag{16}$$

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:

$$\omega_i \;\triangleq\; \underset{n=1,\ldots,N}{\text{argmax}} \|\mathcal{P}_{[\mathbf{\Psi}_{k-i}|\mathbf{\Omega}_i]^{\perp}}\mathbf{x}_n\|, \tag{17}$$

$$r_i \;\triangleq\; \|\mathcal{P}_{[\mathbf{\Psi}_{k-i}|\mathbf{\Omega}_i]^{\perp}}\mathbf{x}_{\omega_i}\|^2, \tag{18}$$

and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-i-1}$ are $k-i-1$ principal left singular vectors of $\mathcal{P}_{\boldsymbol{\Omega}_{i+1}^{\perp}}\mathbf{X}$. Thus, the k columns of $[\boldsymbol{\Psi}_{k-j}|\boldsymbol{\Omega}_j]$, 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 $\boldsymbol{\Omega}_j$ and by SVD-based vectors that best represent (in ℓ_2 sense) the data residuals in the null-subspace of $\boldsymbol{\Omega}_j$. Moreover, each subspace is characterized by it's maximum-norm data representation error r_j. One of these subspaces is to be selected as $\hat{\mathcal{S}}_k$. In the light of our objective to minimize the worst-case representation error, we choose $\hat{\mathcal{S}}_k = \text{range } [\boldsymbol{\Psi}_{k-h}|\boldsymbol{\Omega}_h]$, with the value of h that minimizes the $\ell_{2,\infty}$-norm of residuals, i.e.,

$$h = \underset{j=0,\ldots,k}{\text{argmin}}\ r_j. \tag{19}$$

This policy combines the ℓ_2-based minimization of abundant vector-residual norms with ℓ_∞-based minimization of rare vector residual norms. As we have seen earlier, the rare-vectors have large residuals with respect to principal subspaces found by SVD. This property would cause them to be selected among columns of $\boldsymbol{\Omega}_h$, whereas the abundant vector projections onto the null-space of $\boldsymbol{\Omega}_h$ would lie in the range $\boldsymbol{\Psi}_{k-h}$. A flowchart summarizing the MX-SVD process is shown in Fig. 3.

IV. MX-SVD VS. SVD - SIMULATION RESULTS

In Fig. 4, we show empirical pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, obtained via a Monte-Carlo simulation for $k = r_{abund} + r_{rare} = 5 + 3 = 8$, where r_{abund} is the rank of abundant-vectors subspace and r_{rare} is the number of rare-vectors, which were generated as in the previous example of section II-B by appending orthogonal vectors of equal norms $\{\mathbf{y}_j\}_{j=1}^{r_{rare}}$, $\mathbf{y}_j \in null\ \mathbf{Y}_{abund}^T$. A $10^2 \times 10^5$ matrix $\mathbf{X}$ was generated 1000 times for RSNR $= 10$, $\sigma = 1$. The pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ corresponding to subspace estimation by MX-SVD (dashed line) and SVD (solid line) are shown in Fig. 4(a).

It is clearly seen from the figure that max-norm residuals obtained via MX-SVD have a lower value and have a much narrower pdf, as compared to residuals obtained by SVD. As a matter of fact, the MX-SVD-related pdf is very close to the pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$, which equals to the squared norm of the maximum-norm noise residual. This fact is supported by Fig. 4(b). Here, we plot the empirical pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$ obtained via MX-SVD (dashed line) versus the exact pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$ (solid line), obtained from a model (with the above parameters) that is based on Extreme Value Theory results, presented in Appendix I.

In summary, MX-SVD was designed to yield $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2 \approx \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{Z}\|_{2,\infty}^2$ for $k \geq r$ in the presence of rare vectors, as opposed to SVD, which produces arbitrary large residuals for a range of low RSNR values. The fact that for $k \geq r$ the maximum-norm residuals are governed by the maximum-norm realization

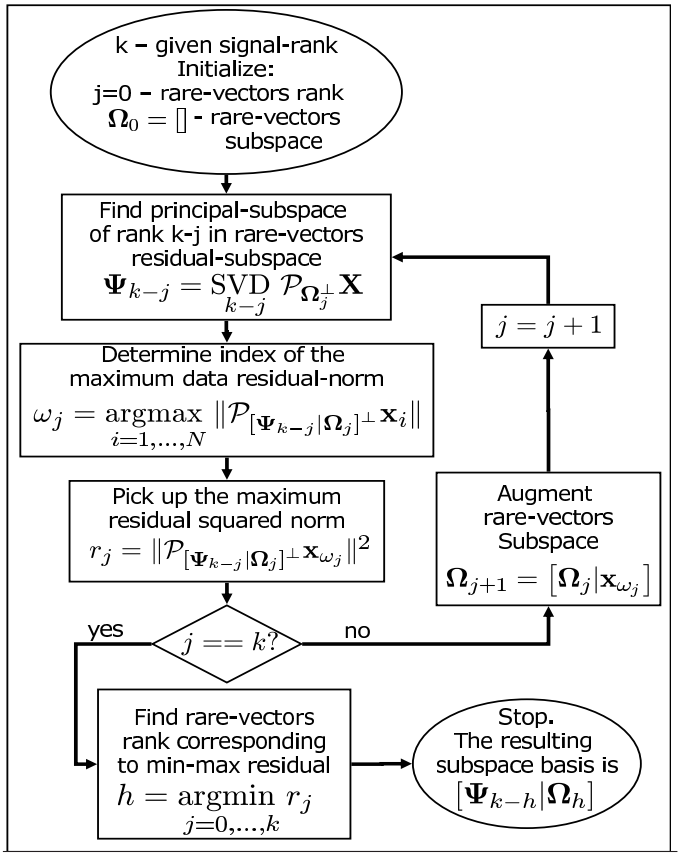


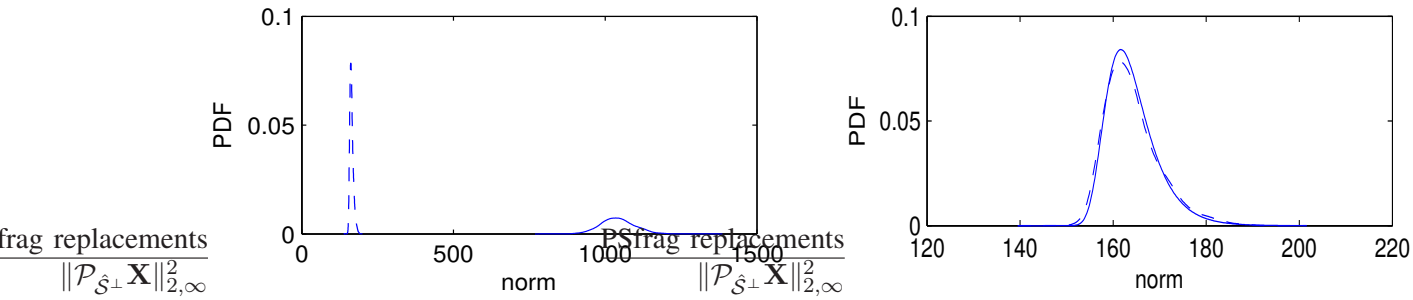
Fig. 3. **MX-SVD flowchart.** For a given signal subspace rank value k, constructs a signal-subspace basis of the form $\hat{\mathcal{S}}_k = [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$, $h \in$ *integers* $[0, k]$, that minimizes $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{X}\|_{2,\infty}^2$, where $\mathbf{\Omega}_h$ is responsible for representing rare-vectors and $\mathbf{\Psi}_{k-h}$ is responsible for representing the remaining (abundant) vectors in the data.

of the noise will be used in the next section for constructing a signal-subspace rank estimator, which is based on statistical properties of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{Z}\|_{2,\infty}^2$.

V. Rank Determination

In this section we construct a signal-subspace rank estimator $\hat{r}$ (recall that the signal-subspace basis may include rare-vectors). This rank estimator is based on examining the maximal data residual norms $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{X}\|_{2,\infty}^2$, for an increasing sequence of k values. The only thing we know about $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{X}\|_{2,\infty}^2$ is that for $k < r$, it could be arbitrarily higher than $\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{Z}\|_{2,\infty}^2$; whereas for $k \geq r$, due the signal-subspace estimation approach, which minimizes $\ell_{2,\infty}$-norm of residuals, one may assume that the maximum-norm data residual is governed by the maximum-norm noise residual, i.e.,

$$\|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{X}\|_{2,\infty}^2 \approx \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{Z}\|_{2,\infty}^2. \tag{20}$$



(a) (b)

Fig. 4. **The pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|^2_{2,\infty}$, obtained via a Monte-Carlo simulation.** (a) The empirical pdfs of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|^2_{2,\infty}$ obtained by MX-SVD (dashed line) and SVD (solid line) for RSNR $= 10$, $\sigma = 1, p = 10^2, N = 10^5$, $k = r_{abund} + r_{rare} = 5 + 3 = 8$ (b) The empirical pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{X}\|^2_{2,\infty}$ by MX-SVD (dashed-line) versus the exact pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}^\perp}\mathbf{Z}\|^2_{2,\infty}$ (solid line).

Guided by (20), we consider a test that determines the rank r as follows in the next section.

A. Signal and noise hypotheses assessment

We assume that for some k, $r_{abund} \leq k \leq r$, the signal-subspace $\hat{\mathcal{S}}_k$, estimated by MX-SVD described above, is close to the subspace of abundant-vectors. This assumption is plausible due to the SVD-part of the MX-SVD process that is designed to represent well the abundant-vectors subspace, which is of rank $r_{abund} \leq r$. As a result, the abundant-vector residuals in the complementary subspace $\hat{\mathcal{S}}_k^\perp$ are governed by the noise contribution, whereas the rare-vector residuals may still include significant signal contributions. Thus, for $k \geq r_{abund}$, the set of all data-vector indices can hypothetically be divided into two subsets according to the properties of data-vector residuals:

$$\begin{aligned} \Gamma_k &\triangleq \{\text{indices } \gamma_j \text{ of abundant-vector residuals}\} \\ \Delta_k &\triangleq \{\text{the remaining data-vector indices } \delta_i \ \}, \end{aligned} \tag{21}$$

such that $j = 1, \ldots, \#\Gamma_k, i = 1, \ldots, \#\Delta_k$ and $\#\Gamma_k \gg \#\Delta_k$, where $\#$ denotes cardinality of a set.

Let η_k be the maximum data-residual squared-norm, $\eta_k = \max\limits_{j=1,\ldots,N} \|\mathcal{P}_{\hat{\mathcal{S}}_k^\perp}\mathbf{x}_j\|^2$. Given η_k, we formulate the following two hypotheses:

$$H_0 \ : \ \eta_k \text{ belongs to } \Gamma_k, \tag{22}$$

$$H_1 \ : \ \eta_k \text{ belongs to } \Delta_k. \tag{23}$$

The following notation will help us to perform a statistical analysis of η_k:

$$\begin{aligned} \nu_k &\triangleq \max_{\gamma_j \in \Gamma_k} \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{x}_{\gamma_j}\|^2 \\ \xi_k &\triangleq \max_{\delta_i \in \Delta_k} \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{x}_{\delta_i}\|^2. \end{aligned} \tag{24}$$

Now, η_k, can be expressed as:

$$\eta_k = \max(\nu_k, \xi_k). \tag{25}$$

Due to the assumption leading to (21), and according to (24), the value of ν_k is governed by the extreme value statistics of maximum-norm noise realizations. Now, we set the rank estimator $\hat{r}$ to be equal to the minimal value of k for which the following condition is satisfied:

$$p(H_0|\eta_k) \geq p(H_1|\eta_k), \tag{26}$$

which means that the optimal rank is reached when there is a higher likelihood that the maximum data-residual squared norm η_k is governed by the noise statistics (i.e., it doesn't include significant signal contributions).

In order to evaluate the conditional probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$, one has to specify pdfs $f_{\nu_k}(\cdot)$ and $f_{\xi_k}(\cdot)$, or, equivalently, cdfs $F_{\nu_k}(\cdot)$ and $F_{\xi_k}(\cdot)$. Whilst the probability of maximum-norm noise realization ν_k can be determined by Extreme Value Theory results, as shown in Appendix I, the pdf of ξ_k is generally unknown. The only thing we know about ξ_k is that at each iteration k, it has to be less or equal η_{k-1}. A possible choice for $f_{\xi_k}(\cdot)$ is therefore,

$$\xi_k \sim U[0, \eta_{k-1}], \tag{27}$$

where U denotes a uniform distribution.

Now, it can be shown (see Appendix II for details) that posterior hypotheses probabilities are given by:

$$p(H_0|\eta_k) = \frac{\eta_k f_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k) + F_{\nu_k}(\eta_k)}, \tag{28}$$

$$p(H_1|\eta_k) = \frac{F_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k) + F_{\nu_k}(\eta_k)}, \tag{29}$$

where the expressions above are valid for $0 \leq \eta_k \leq \eta_{k-1}$. It is important to note, however, that the functional form of the posterior conditional probabilities, as given in (28) and (29), does not depend on η_{k-1}. Moreover, due to a successive application of MX-SVD for an increasing sequence of k values, it

is guaranteed that $0 \leq \eta_k \leq \eta_{k-1}$. Therefore, in the forthcoming expressions, we omit explicit mention of the argument boundaries.

Fig. 5(a) shows the corresponding graphs of these posterior probabilities for a residual-subspace rank $l = p - k = 10^2$, where p is the dimensionality of the data vectors $\mathbf{x}$, the total number of data vectors $N = 10^5$, and the noise std $\sigma = 1$. It is clearly seen that the transition region between hypotheses is steep and narrow. Actually, its width depends on the form of f_{ν_k} (see Fig. 5(b)), which is well-localized, as explained in Appendix I.

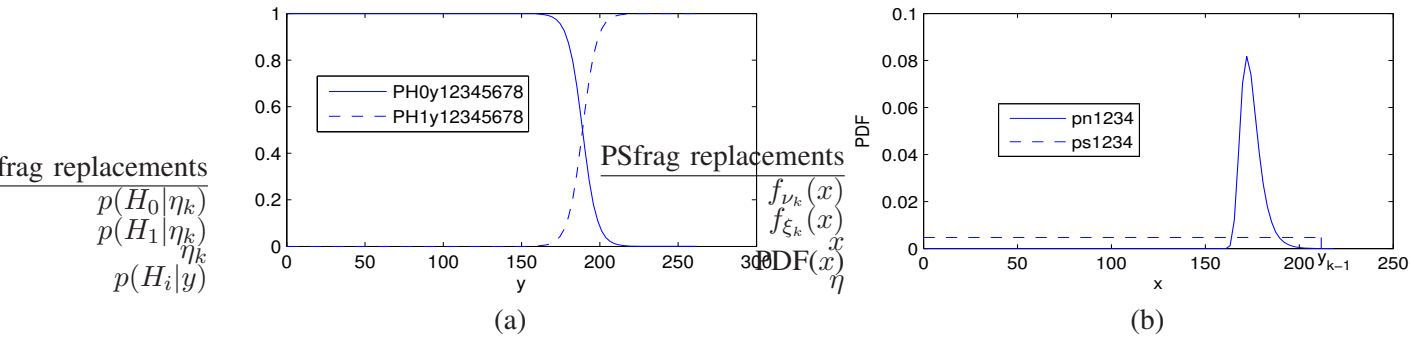


Fig. 5. **a) posterior conditional hypotheses probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$ b) distributions of maximum squared-norm of rare (solid line) and abundant (dashed line) vector residuals.** For residual-subspace rank $l = 10^2$, total number of data vectors $N = 10^5$, and the noise std $\sigma = 1$.

In summary, the signal-subspace rank is determined by applying MX-SVD and examining condition (26) for an increasing sequence of k values. As the maximum-norm residual becomes low and (26) becomes true, it can no longer be confidently associated with the signal contribution, and the procedure is terminated. The estimated rank is equal to the last-examined k value. As was already noted above, this combination of applying MX-SVD and examining condition (26) for an increasing sequence of k values, defines what we called Maximum Orthogonal Complement Algorithm (MOCA), and is summarized next.

B. MOCA summary for combined subspace and rank determination

In this subsection we summarize the proposed approach of signal-subspace and rank determination via MOCA by presenting its major parts in the flowchart of Fig. 6.

The algorithm begins with an initial guess for the signal-subspace rank, such as $k = 1$. At each rank value iteration, the signal-subspace basis $\mathbf{\Phi}_k = [\mathbf{\Psi}_{k-h}|\mathbf{\Omega}_h]$ is obtained via MX-SVD of section III, using the conjectured rank k. Then the data maximum residual-norm is calculated in the null space of $\mathbf{\Phi}_k$. This norm is tested in order to decide if it belongs to the noise hypothesis (this decision is performed by

evaluating inequality (26)). If the noise hypothesis passes, the algorithm is terminated, and the estimated signal-subspace and rank equals to the span of the last obtained $\mathbf{\Phi}_k$, and to the last value of k, respectively. Otherwise, the rank conjecture k is incremented and a new iteration is carried out.

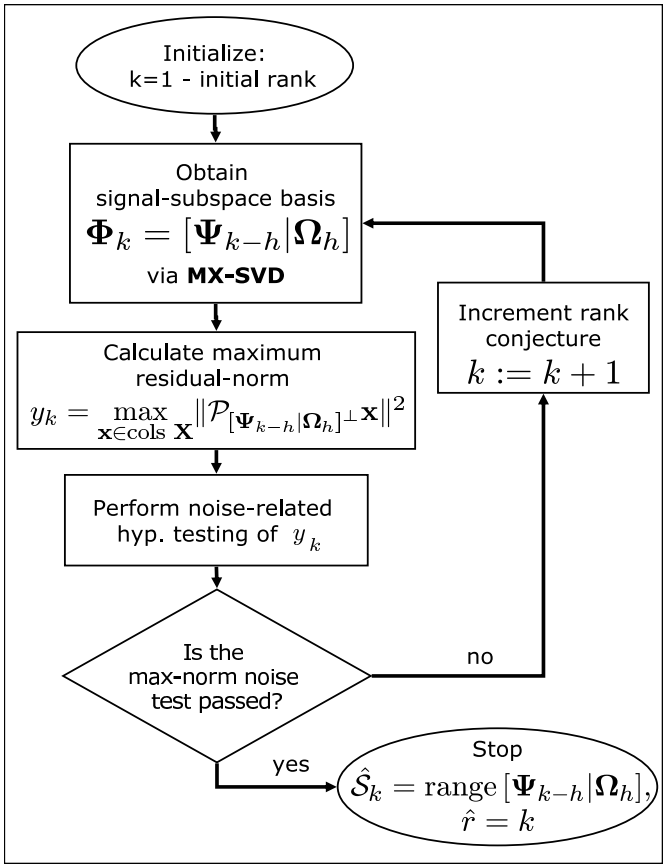


Fig. 6. **Maximum Orthogonal Complement Algorithm (MOCA) flowchart.**

VI. Comparison of rank determination by MOCA vs. MDL

In this section we compare the performance of MOCA with that of the Minimum Description Length (MDL) approach for signal-subspace rank determination.

A. MDL basics

MDL is a widely-used model-order determination criterion, based on coding arguments and the minimum description length principles [12],[13]. The same rule has been also obtained via a rather different approach, based on a Bayesian Information Criterion (BIC) [14]. Thus, in [11] it is proposed to apply the MDL for determining the model-order of (1), with $\{\mathbf{s}_i\}$ being an ergodic Gaussian process with a positive definite covariance matrix and the noise variance σ^2 is unknown.

The MDL was also proven in [11] to be consistent in terms of yielding the true signal-subspace rank, with probability one, as the sample size N increases. It is based on minimizing the following criterion with respect to k:

$$\text{MDL}(k) = -\ln f(\mathbf{X}|\hat{\mathbf{\Theta}}(k)) + \frac{1}{2}\eta \log N, \tag{30}$$

where $f(\cdot)$ is a family of probability densities parameterized by $\mathbf{\Theta}(k)$, and η denotes the number of model degrees of freedom. In our case, where σ^2 is known, manipulation of the results in [11] gives:

$$\text{MDL}(k) = \sum_{i=1}^{k} \log(\hat{l}_i) + (p-k)\log(\sigma^2) + k + \sum_{i=k+1}^{p} \frac{\hat{l}_i}{\sigma^2} + k(2p-k)\frac{\log(N)}{N}, \tag{31}$$

where $\{\hat{l}_i\}_1^p$ denote eigenvectors of the data-covariance matrix $\mathbf{R} \triangleq E\{\mathbf{x}\mathbf{x}^T\}$, and σ^2 is the known noise variance.

B. Simulation of rank determination by MOCA vs. MDL

In this subsection we compare the results of applying MOCA and MDL to simulated examples, in the presence of rare vectors, and assess their performance in terms of rank errors expressed by rank-RMSE defined by $e_{rank} \triangleq \sqrt{E(r-\hat{r})^2}$.

Fig. 7 shows the performance of MOCA vs. MDL for $r = r_{abund} + r_{rare} = 5 + 10 = 15$, SNR $= 100$ (the rare vectors were generated as in the example of section II-B); with Fig. 7 (a) and (b) corresponding to different sizes of $N = 10^4$ and $N = 10^5$, respectively. MOCA and MDL were tested 50 times for each value of RSNR . Then, the rank-determination errors were calculated and plotted. The error e_{rank} obtained by MDL for a range of low RSNR values, which is a function of N, is equal to 10 (the rare-vectors subspace rank r_{rare}). In other words, the MDL completely fails to determine r at low RSNR values. The dependence of e_{rank} on N is obvious - the larger the sample size N is, the more "blind" becomes MDL to rare-vectors, which have to be much stronger in order to become apparent to MDL. Thus, the correct rank determination by MDL starts only at very high values of RSNR. In contrast to MDL, MOCA performs much better, with low values of e_{rank} obtained already at a very low RSNR value.

It turns out, that the probability of rank determination error by MOCA becomes small and approximately constant already for RSNR as small as 2 (see Appendix III for details). This turning point is marked by a heavy dot-dashed vertical line in Fig. 7.

It is important to note that the simulations above were designed to reflect a typical situation seen in hyperspectral images, in which the background process is characterized by $SNR \approx 100$ (20dB), while anomalies are characterized by $RSNR \leq 30$. Hence, the simulations above indicate that MDL is expected to be "blind" to the anomaly subspace rank in typical hyperspectral images, whereas MOCA is expected to succeed in estimating the rank.

A reasonable question that arises is how to identify the transition point, below which one should use MOCA due to its ability to recover the rank at low RSNR values, and above which one could use MDL due to its computational simplicity. We turn to (31) and notice that MDL(k) has to accept its minimum at k. That means that the increase in penalty (the last term of (31)) has to be smaller than the decrease in minus log-likelihood (the first part of (31)) in the transition from $k-1$ to k and, respectively, larger in the transition from k to $k+1$. Now, due to construction of $\mathbf{Y}$ in simulations (see II-B)), the eigenvalue $\hat{l}_k$ stemming from rare vectors is assumed to satisfy:

$$\hat{l}_k \approx \sigma^2 + \|\mathbf{y}_{rare}\|^2/N = \sigma^2 + RSNR(p-k)/N. \tag{32}$$

By neglecting k with respect to p (since $k \ll p$) and approximating $\hat{l}_i \approx \sigma^2$ for $i > k$, then, with some straightforward manipulations, one obtains that the equilibrium between the change in penalty and change in log-likelihood (when k is changed to $k+1$) is reached when:

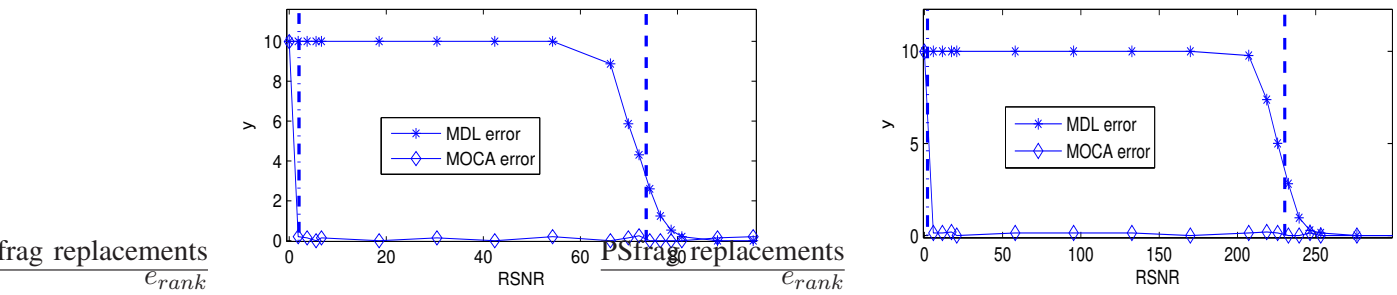
$$-\log(\sigma^2 + RSNR\frac{p}{N}) + RSNR\frac{p}{N\sigma^2} = 2p\frac{\log(N)}{N}. \tag{33}$$

By numerically solving (33) with respect to RSNR, one obtains the turning point in RSNR value below which the MDL is expected to be unreliable in determining contribution of rare-vector to rank. This turning point is marked by a heavy dashed vertical line in Fig. 7.

C. Comparing MOCA with MDL on real data

In this section we compare results of MOCA and MDL for signal-subspace and rank determination of hyperspectral images. We then compare MOCA and MDL-SVD performances in dimensionality reduction of hyperspectral images by applying MDL and MOCA on a bank of about 50 hyperspectral cubes of size 400×450 with 65 spectral bands. Due to space limitations, results for a typical cube are demonstrated here.

One of hypespectral bands of this cube is shown in Fig. 8. Each pixel in this hyperspectral image corresponds to a 65×1 vector. MOCA assumes the noise to be statistically independent between spectral



(a) (b)

Fig. 7. **MOCA vs MDL comparison via Monte Carlo simulations.** The rank estimation error $e_{rank} = \sqrt{E(r - \hat{r})^2}$ in the presence of 10 rare-vectors as a function of RSNR, for (a) $N = 10^4$, (b) $N = 10^5$. The heavy dashed and dot-dashed vertical lines delimit a region in which MOCA is reliable enough and has better performance than MDL.

bands. Therefore, in order to make the noise i.i.d., the noise std in each band was estimated and normalized to 1 by scaling the data.

First, MOCA was applied on the upper part of the image shown in Fig. 8 that is delimited by horizontal and vertical white lines. According to ground-truth evidence, this part corresponds to a "pure background signal" stemming from agricultural fields radiance. Indeed, the signal-subspace determined by MOCA is given by $\hat{\mathcal{S}}_{\text{I}} = \mathbf{\Psi}_7$, which corresponds to $k = 7, h = 0$; i.e., no rare-vectors were selected in order to represent best the signal-subspace in this subimage. Then, MOCA was applied on the entire image producing $\hat{\mathcal{S}}_{\text{II}} = [\mathbf{\Psi}_6 | \mathbf{\Omega}_4]$, which corresponds to $k = 10, h = 4$. Such a result can be explained by the presence of anomaly pixels (marked by circles) located at the bottom of the image. According to the ground truth, these pixels belong to vehicles, which are anomalous to the natural surroundings in the image. Thus, there are 4 data vector pixels comprising $\mathbf{\Omega}_4$ columns, which represent the anomaly pixels subspace in the data.

It should be stressed that the number of columns in $\mathbf{\Omega}_4$ may be less than the number of anomaly pixels in the data, since the columns of $\mathbf{\Omega}_4$ are intended to span the anomaly pixels subspace, which may be of a rank lower than the number of anomaly pixels. Moreover, since the rare-vector subspace and the background-subspace are not orthogonal to each other, the columns of $\mathbf{\Omega}_4$ may span a subspace close to the background subspace $\mathbf{\Psi}_7$, found initially in $\hat{\mathcal{S}}_{\text{I}}$, which may produce a $\ell_{2,\infty}$-norm of residuals small enough in order to stop at an earlier MOCA iteration. This explains why the background subspace rank is lower in $\hat{\mathcal{S}}_{\text{II}}$ than in $\hat{\mathcal{S}}_{\text{I}}$ (the pure-background case with no anomalies).

Turning to the examination of MDL performance, we note that MDL is known to be sensitive to deviations from the white noise assumption [34]. We have found that the noise normalization preprocessing

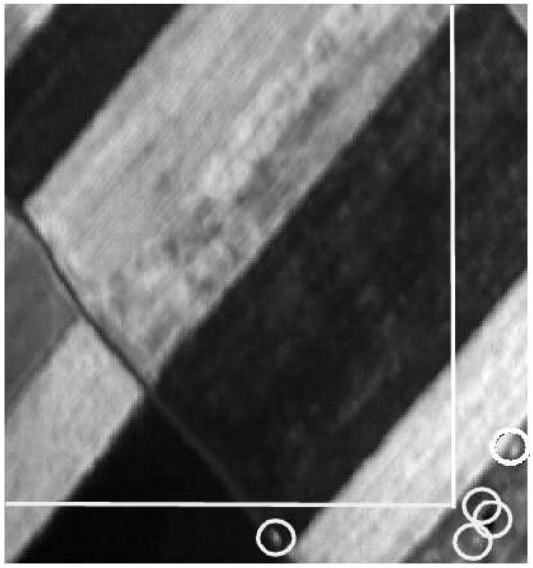


Fig. 8. **Signal-subspace and rank determination in a hyperspectral image.** MOCA was applied on (i) the subimage above the white lines produces $\hat{\mathcal{S}}_{\rm I} = \mathbf{\Psi}_{\rm I}$, (ii) the entire image includes anomalies marked by circles, producing $\hat{\mathcal{S}}_{\rm II} = [\mathbf{\Psi}_{\rm II}|\mathbf{\Omega}_{\rm II}]$. The MDL-estimated rank in both cases is 7.

that produced good results with MOCA, isn't sufficient for a proper operation of MDL, since it doesn't compensate for small correlations between noise components in adjacent bands due to a crosstalk between adjacent sensors, and still leaves noise component variances different. We have applied, therefore, the Robust MDL (RMDL) algorithm of [34] (which assumes different diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$), but with a slight modification, to account for correlations between the adjacent noise components. The modification we applied to RMDL is described in Appendix IV.

We have applied the modified RMDL algorithm for rank estimation on the above mentioned hyperspectral images: the pure-background subimage and the anomaly-containing entire image. In the pure-background subimage case, the MDL has produced a rank of 7, which is in accordance with the result of MOCA. However, in the case of the entire image, which contains rare vectors, the MDL algorithm misses the contribution of rare-vectors to signal-subspace rank, leaving the rank value at 7, whereas MOCA manages to detect the contribution of anomalies to the signal-subspace and rank producing a higher rank of 10 corresponding to both the background and rare-vector pixels.

Now, all hyperspectal pixels were projected onto the subspace found by SVD, of rank found by MDL, as well as onto the signal-subspace basis $\hat{\mathcal{S}}_{\rm II}$ found by MOCA. In Fig. 9 we show squared norms of residuals corresponding to (a) MDL-SVD, and (b) MOCA based subspaces. It is clearly seen that MOCA-based dimensionality reduction better represents all pixels in the image including the anomalies,

compared to MDL-SVD based dimensionality reduction, which misrepresents anomaly pixels producing high-intensity residuals (white blobs in Fig. 9 (a)) at their location.

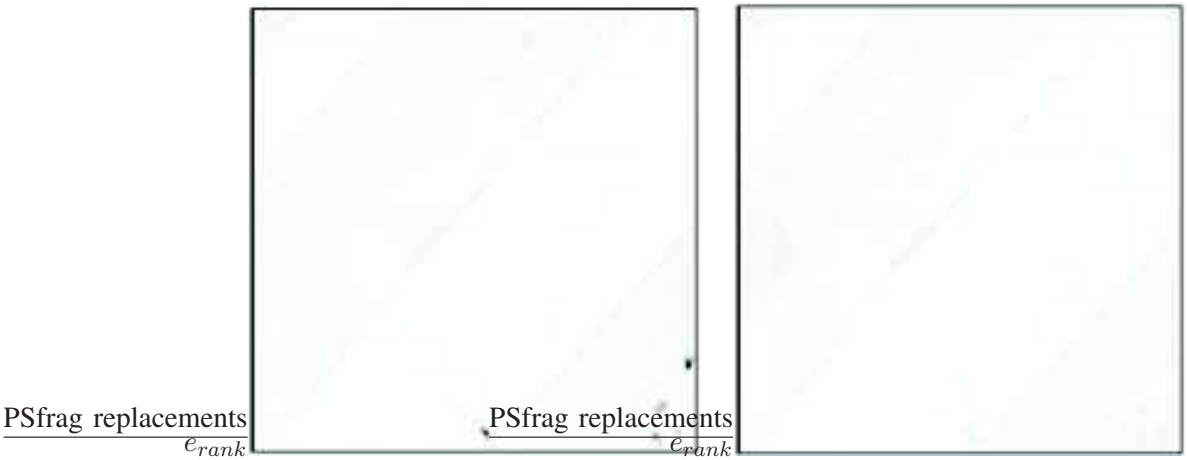


Fig. 9. **Squared norms of residuals corresponding to (a) MDL-SVD, and (b) MOCA based subspaces.**

VII. Conclusion

In conclusion, in this work we have proposed an algorithm for redundancy reduction of high-dimensional noisy signals, named MOCA, which is designed for applications where a good representation of *both* the abundant and the rare vectors is essential. The combined subspace of rare and abundant vectors is obtained by using the proposed $\ell_{2,\infty}$-norm that penalizes individual data-vector miss-representations. Since this criterion is hard to optimize, a sub-optimal greedy algorithm is proposed. It uses a combination of SVD and direct selection of vectors from the data to form the signal-subspace basis. The rank is determined by applying Extreme Value Theory results to model the distribution of the maximal noise-residual ℓ_2-norms. In simulations, conducted for various rare-vectors signal-to-noise conditions, the proposed approach is shown to yield good results for practically-significant RSNR values (RSNR essentially measures the SNR of rare-vectors with respect to noise), for which the classical methods of SVD and MDL fail to determine correctly the signal-subspace and rank, respectively, of high dimensional signals composed of abundant and rare vectors.

The proposed approach was also applied for the signal-subspace and rank determination of a hyper-spectral image with and without anomaly pixels. The results of MOCA were found to be equal to those of MDL (or when necessary RMDL) for the pure-background subimage, whereas in the presence of anomalies, MOCA has detected a higher rank than MDL, while MDL produced the same rank as in the

pure-background case. This indicates that MDL failed to determine correctly the signal-subspace rank of a hyperspectral image composed of both abundant and rare vectors, whereas MOCA succeeded in representing it well.

The proposed approach can be further developed for anomaly detection and classification, which is an objective of our current activity.

APPENDIX I
DISTRIBUTION OF MAXIMUM-NORM NOISE REALIZATIONS

In this appendix we characterize the pdf $f_{\nu_k}(\cdot)$ of section V-A. We assume that the noise is a zero-mean white Gaussian process, with known standard deviation σ. Then, its residual squared norms

$$\zeta_{k,i} \triangleq \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}} \mathbf{z}_i\|^2, \tag{34}$$

$i = 1, \ldots, N$, have a Chi-squared distribution of order $l \triangleq$ rank $\hat{\mathcal{S}}_k^{\perp} = p - k$, denoted by $\chi^2(l, \sigma^2)$ with the following pdf [30]:

$$f(u) = \frac{1}{2^{l/2}\Gamma(l/2)\sigma^2} \left(\frac{u}{\sigma^2}\right)^{(l/2)-1} e^{-u/2\sigma^2}. \tag{35}$$

For large l, the Central Limit Theorem can be used to obtain the following approximation:

$$\zeta_{k,i} \sim \chi^2(l, \sigma^2) \approx \mathcal{N}\left(l\sigma^2, 2l\sigma^4\right). \tag{36}$$

Now, the limiting distribution of ν_k, which satisfies

$$\nu_k = \max_{i=1,\ldots,N} \zeta_{k,i}, \tag{37}$$

can be obtained using the following Extreme Value Theory result:

Theorem 1: [27]

If $\{\zeta_i\}_{i=1}^N$ is i.i.d., with absolutely continuous distribution $F(x)$ and density $f(x)$, and letting

(i) $h(x) = f(x)/(1 - F(x))$

(ii) $b_N = F^{-1}(1 - \frac{1}{N})$

(iii) $a_N = h(b_N)$

(vi) $\omega = lim_{x \to x^*} \frac{dh(x)}{dx}$,

where x^* is the upper end-point of F,

then, for $M_N = \max\{\zeta_1 \dots \zeta_N\}$,

$$P(a_N(M_N - b_N) \leq u) \underset{N\to\infty}{\longrightarrow} \begin{cases} \exp(-e^{-u}), & if\ \omega = \infty \\ \exp\{-[1+\frac{u}{\omega}]^{\omega}\}, if\ \omega < \infty \end{cases}, \tag{38}$$

The proof is found in [27].

In words: Theorem 1 says that the maximum of N i.i.d random variables has a limiting distribution that depends on ω - a parameter derived from their individual distributions. For the purposes of the present work, we consider normal and chi-squared distributions, which lead to $\omega = \infty$.

Therefore, from (38), the limiting distribution of interest is

$$\mathcal{G}(u) \triangleq \exp(-e^{-u}), \tag{39}$$

also known as the Gumbel distribution [1]. The mean and std of a variable distributed as (39) are $\eta = 0.5772$ and $\gamma = 1.6450$, respectively. The normalizing coefficients a_N and b_N are also functions of the ζ_i distribution. Theorem 1 also describes how to calculate the normalizing coefficients given the distribution function of ζ_i.

Unfortunately, there are no known analytical expressions for the normalizing coefficients a_N and b_N corresponding to $\{\zeta_{k,i}\}$ (defined in (34)) that are chi-square distributed. In our evaluations of the asymptotic pdf of ν_k in Fig. 4(b) above and in the sequel, we used the results of Theorem 1 to calculate a_N and b_N numerically. Note, that a_N and b_N are also functions of l and σ, since they depend on the $\chi^2(l, \sigma^2)$ distribution, which is a function of l and σ.

However, in order to explain why the pdf of $\|\mathcal{P}_{\hat{\mathcal{S}}^{\perp}}\mathbf{Z}\|^2_{2,\infty}$, shown in Fig. 4(a) and Fig. 5(b), is so narrow; one can use the approximation in (36) to obtain the following asymptotic analysis, which can be conducted analytically. It can be shown [25] that for $\{\zeta_i\}$ of Theorem 1, which are Gaussian, M_N is distributed as follows:

$$P(M_N \leq u) \underset{N\to\infty}{\longrightarrow} \mathcal{G}(a_N(u - b_N)), \tag{40}$$

[1]Extreme Value Distributions are the limiting distributions of the minimum or the maximum of a very large collection of random observations from the same arbitrary distribution. Gumbel (1958), [28] showed that for any well-behaved initial distribution (i.e., F(x) is continuous and has an inverse), only a few models of limiting distributions are needed, depending on whether one is interested in the maximum or the minimum, and also if the observations are bounded from above or below (see [25]).

with

$$\begin{aligned} a_N &= (2\ln N)^{1/2} \\ b_N &= (2\ln N)^{1/2} - \\ &\quad \frac{1}{2}(2\ln N)^{-1/2}(\ln\ln N + \ln 4\pi). \end{aligned}$$

Therefore,

$$P(\nu_k \leq x) \approx \mathcal{G}\left(a_N\left[\frac{x-\sigma^2 l}{\sigma^2\sqrt{2l}} - b_N\right]\right) \tag{41}$$

with mean and std:

$$\mu_N = \sigma^2\left(\frac{\sqrt{2l}}{a_N}\eta + b_N\sqrt{2l} + l\right) \tag{42}$$

$$\sigma_N = \frac{\sigma^2\sqrt{2l}}{a_N}\gamma \tag{43}$$

While this approximation doesn't provide us with an accurate mean and std of ν_k, it is instructive to look at the following ratio that defines a relative width of a pdf for $N \gg 1$, $l \gg 1$:

$$\frac{\mu_N}{\sigma_N} \propto 2\ln N + \sqrt{l\ln N}. \tag{44}$$

It is observed that this ratio doesn't depend on σ^2, and it is log-dependent on N. Thus, the ratio μ_N/σ_N tends to infinity as $N \to \infty$ or $l \to \infty$. For example, for $l = 100, N = 10^5$ and white noise, $\mu_N/\sigma_N \approx 23$ corresponding to quite a small relative width. The dominant factor in obtaining such a high ratio is the high dimensionality of $l = 100$.

APPENDIX II
DERIVATION OF POSTERIOR HYPOTHESIS PROBABILITIES

In the following, we derive the conditional probabilities $p(H_0|\eta_k)$ and $p(H_1|\eta_k)$ in (28) and (29), based on pdfs f_{ν_k} and f_{ξ_k}:

$$\begin{aligned}
f(H_0,\eta_k) &= f_{\nu_k}(y)p(\xi_k<\eta_k) = \\
&\quad f_{\nu_k}(\eta_k)F_{\xi_k}(\eta_k) = f_{\nu_k}(\eta_k)\frac{\eta_k}{\eta_{k-1}}, \\
f(H_1,\eta_k) &= f_{\xi_k}(\eta_k)p(\nu_k<\eta_k) = \\
&\quad f_{\xi_k}(\eta_k)F_{\nu_k}(\eta_k) = F_{\nu_k}(\eta_k)\frac{1}{\eta_{k-1}}, \\
f_{\eta_k}(\eta_k) &= f(H_0,\eta_k)+f(H_1,\eta_k) = \\
&\quad \frac{1}{\eta_{k-1}}\left[\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)\right], \\
p(H_0|\eta_k) &= \frac{f_{\nu_k}(\eta_k)F_{\xi_k}(\eta_k)}{f_{\eta_k}(\eta_k)} = \frac{\eta_k f_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)}, \\
p(H_1|\eta_k) &= \frac{f_{\xi_k}(\eta_k)F_{\nu_k}(\eta_k)}{f_{\eta_k}(\eta_k)} = \frac{F_{\nu_k}(\eta_k)}{\eta_k f_{\nu_k}(\eta_k)+F_{\nu_k}(\eta_k)},
\end{aligned}$$

which are the expressions shown in (28) and (29).

APPENDIX III
ASSESSMENT OF MOCA RELIABILITY IN TERMS OF RSNR

In the following we assess the dependence of MOCA rank estimation error on the value of RSNR.

Let's recall that the RSNR notion was introduced in the context of SVD performance assessment in the presence of rare-vectors. It measures the ratio between the contribution of rare-vectors and the contribution of noise to the signal covariance matrix. Thus, being ℓ_2-based, RSNR is an ambiguous measure for MOCA performance assessment, which is affected by individual data-vector contributions. For example, two identical rare-vectors of the same ℓ_2-norm value l, have the same RSNR as that of a single rare-vector of an ℓ_2-norm value of $l\sqrt{2}$, and thus have the same SVD performance. However, MOCA may behave differently in each of the two cases in this example. Thus, in some applications, there is typically only one rare-vector out of 10^5 data-vectors, whereas in other applications, even 10 collinear vectors out of 10^5 are considered to be rare. Different rare-vector multiplicities cause MOCA to depend differently on the RSNR. In order to eliminate this ambiguity, we constrain the rare-vectors in the following analysis to be linearly independent. Otherwise, the RSNR value should be corrected by

an appropriate rare-vectors multiplicity factor in order to obtain an equivalent MOCA performance.

If the SNR of abundant vectors is high enough, then we can assume that for $k \geq r_a$, where r_a is the abundant vectors subspace rank, the SVD-part of MOCA estimates well the abundant vectors subspace, and that MOCA iterations don't terminate before $k = r_a$. Thus, in the complementary subspace for $r_a \leq k < r$, one would find only residuals of abundant vectors, composed of noise only, and residuals of rare-vectors. Let's denote

$$\tilde{\mathbf{Y}}_{rare} \triangleq \mathcal{P}_{\mathbf{Y}_{abund}^{\perp}} \mathbf{Y}_{rare}, \tag{45}$$

i.e., the projection of the rare-vectors sub-matrix onto the abundant-vectors null-space. Our purpose here is to characterize RSNR values for k values satisfying $r_a \leq k \leq r$, for which there is a high probability that rare-vectors will be selected among $\mathbf{\Omega}_{k-r_a}$ columns (see (15)).

Let's assume that for some iteration k, $r_a < k \leq r$, the matrix $\mathbf{\Omega}_{k-r_a}$ is composed of rare vectors. We are looking for conditions on RSNR that guarantee selecting the next rare-vector at iteration k as in (15), with probability close to 1. This RSNR value would also justify the assumption on the matrix $\mathbf{\Omega}_{k-r_a}$ above, since (as we'll see later) it would guarantee the rare-vectors selection for all $r_a < k \leq r$, with probability close to 1. If one neglects the effect of noise on the rare vectors selected in $\mathbf{\Omega}_{k-r_a}$, then the ℓ_2-norms of the remaining $r-k$ rare vectors can equivalently be obtained as the last $r-k$ diagonal entries of the upper triangular matrix $\mathbf{R}$ obtained via the following QR decomposition:

$$\mathbf{QR} = \tilde{\mathbf{Y}}_{rare}\mathbf{\Pi}, \tag{46}$$

where $\mathbf{\Pi}$ is a permutation matrix that moves $\tilde{\mathbf{Y}}_{rare}$ columns of rare-vectors selected in $\mathbf{\Omega}_{k-r_a}$ to the leading positions. Now, we use the following lemma in order to obtain a relation between the RSNR of $\mathbf{Y}$ and the diagonal entries of $\mathbf{R}$.

Lemma 1: The minimal singular value s_{min} of a full-rank $m \times n$ matrix $\mathbf{M}$ with $m > n$, satisfies $s_{min} \leq \rho_j,\ j = 1, \ldots n$, where ρ_j are the diagonal entries of a triangular matrix in the QR decomposition of $\mathbf{M\Pi}$, with $\mathbf{\Pi}$ - any permutation matrix.

Proof: Let ρ_j be a diagonal entry for some $j = 1, \ldots, n$, and let $\hat{\mathbf{\Pi}}$ be another permutation matrix that moves column j of $\mathbf{M\Pi}$ to the last. Then, the corresponding $\hat{\rho}_n$ of $\mathbf{M\Pi\hat{\Pi}}$ satisfies: $\hat{\rho}_n \leq \rho_j$, since it is a norm of a projection onto a smaller (contained) subspace. Now, according to [2], the following holds: $\hat{s}_{min} \leq \rho_n$. Therefore, $s_{min} \leq \rho_j$.

Using the lemma above and the definition of RSNR (7), one obtains:

$$\varsigma_k \geq \text{RSNR}\sigma^2(p-r), \tag{47}$$

where $\varsigma_k \triangleq \|\tilde{\mathbf{y}}_{max}\|^2$, and $\tilde{\mathbf{y}}_{max}$ is the maximum-norm rare-vector residual in $\hat{\mathcal{S}}_k^{\perp}$. Since the termination condition of MOCA is based on testing the maximum squared norm of residuals $\eta_k = \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}\|_{2,\infty}^2$, it is important to calculate the pdf of η_k, which satisfies:

$$\eta_k = \quad \max\ (\xi_k, \nu_k), \tag{48}$$

where,

$$\nu_k = \quad \|\mathcal{P}_{\hat{\mathcal{S}}_k^{\perp}}\mathbf{X}_{abund}\|_{2,\infty}^2 \tag{49}$$

$$\xi_k = \quad \|\tilde{\mathbf{y}}_{max} + \mathbf{n}\|^2, \tag{50}$$

we also assume here that the RSNR value is large enough, so that:

$$\underset{\tilde{\mathbf{y}}_i \in \text{columns } \tilde{\mathbf{Y}}}{\text{argmax}} \|\tilde{\mathbf{y}}_i + \mathbf{n}\| = \underset{\tilde{\mathbf{y}}_i \in \text{columns } \tilde{\mathbf{Y}}}{\text{argmax}} \|\tilde{\mathbf{y}}_i\|, \tag{51}$$

with probability close to 1.

Now, the distribution function of η_k for $r_a \leq k < r$ is given by:

$$F_{\eta_k}(\cdot) = \mathcal{G}_{p-k}(\cdot) NC\chi^2_{p-k,\delta}(\cdot), \tag{52}$$

where $\mathcal{G}_{p-k}(\cdot)$ is the Gumbel distribution of the noise max-norm with $p-k$ degrees of freedom, as described in Appendix I, and $NC\chi^2_{p-k,\delta}(\cdot)$ is the noncentral chi-square distribution [32], with $p-k$ degrees of freedom and δ is its non-centrality parameter. The results of [32] and relation (47) can be used to obtain:

$$\delta = \frac{\varsigma_k}{\sigma^2} \geq \text{RSNR}\ (p-r). \tag{53}$$

The pdf of η_{r-1}, $f_{\eta_{r-1}}$, corresponding to a situation where $\varsigma_{r-1} = \text{RSNR}\sigma^2(p-r)$ (selecting the worst case in (47)), RSNR $= 2$, $p = 100$, $r = 10$, $r_a = 5$, $\sigma = 1$, $N = 10^4$ is shown in Fig. 10, solid line. The choice of $k = r-1$ is arbitrary for numerical demonstration purpose only. Now, the distribution of η_r, $F_{\eta_r}(\cdot)$, equals to the distribution of maximum-norm noise residual $\mathcal{G}_{p-r}(\cdot)$, since $\hat{\mathcal{S}}_r^{\perp}$ is supposed to include only noise. The pdf of η_r, f_{η_r}, is plotted in dashed line. The rank-determination threshold τ_{r-1}

at iteration $r-1$ (marked by a vertical line) equals to η_{r-1}, satisfying:

$$p(H_0|\eta_{r-1}) = p(H_1|\eta_{r-1}), \tag{54}$$

where H_0, H_1 are defined in subsection V-A.

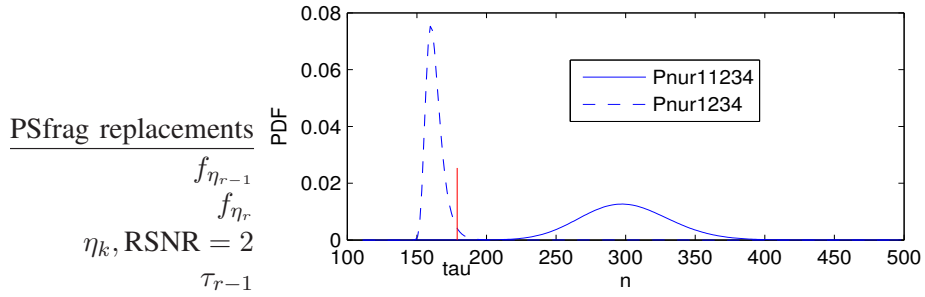


Fig. 10. Pdf of the maximum residual norm η_{r-1} and η_r for $k = r-1$, $\varsigma_{r-1} = \text{RSNR}\sigma^2(p-r)$, RSNR $= 2$, $p = 100$, $r = 10$, $r_a = 5$, $\sigma = 1$, $N = 10^4$ at iteration $r-1$ (solid line) and iteration r (dashed line), respectively. The rank-determination threshold τ_r at iteration r is marked by a vertical line.

Now, the probability p_u of rank underestimation, given that iteration $r-1$ is reached, is given by $p_u = F_{\eta_{r-1}}(\tau_{r-1})$, which for the parameters above is of the order of 10^{-6}! It turns out that for the parameters above, the order of the rank underestimation error is approximately the same for all k values $r_a \leq k < r$, which is small enough to be neglected.

It is important to note that typically, $f_{\eta_{r-1}}$ would lie farther from the threshold τ_{r-1}, since selecting equality in (47), in this example, corresponds to the worst case. This decreases the probability of the rank underestimation even further. Due to properties of $\mathcal{G}_{p-k}(\cdot)$, the distribution of η_k has a weak $\log N$ dependence on the data sample size N (see (41)). Whereas $NC\chi^2_{p-k,\delta}(\cdot)$ doesn't depend on N at all. Therefore, the rank underestimation error is also negligible for $N = 10^3$ as well as for $N = 10^5$.

The probability of rank overestimation p_o at iteration $k = r$, is given by $p_u = 1 - F_{\eta_r}(\tau_r) = 1 - \mathcal{G}_{p-r}(\tau_r)$, which for the parameter values above gives $p_o \approx 0.027$. This value is nearly constant for all RSNR values above 2, which, as we have seen earlier, guarantee a negligible p_u. It can be decreased by modifying the hypotheses equality test of (54) to the following likelihood ratio test:

$$p(H_0|\eta_r) \leq \gamma p(H_1|\eta_r), \qquad \gamma < 1. \tag{55}$$

This should produce a lower error-rate at the expense of a higher τ_{r-1}. Fortunately, as it is clearly seen in Fig. 10, the pdf $f_{\eta_{r-1}}$ lies far from τ_{r-1}, which means that a lower p_o can be obtained by choosing an appropriate $\gamma < 1$ leaving p_u still negligible.

APPENDIX IV
ROBUST MDL WITH A MODIFICATION THAT ACCOUNTS FOR NOISE DEPENDENCE BETWEEN BANDS

In section VI-C we apply the RMDL approach [34] as an ℓ_2-based alternative to the classical MDL approach for signal-subspace rank determination. The assumption of RSNR that the noise covariance matrix is diagonal, but with different diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$, makes the algorithm robust to deviations of noise variances from being equal in all spectral bands. In order to model also the observed small dependence of noise components between adjacent bands, we assume that the secondary-diagonal noise covariance matrix entries are all-equal to a parameter β_k. As in [34], let's define $\sigma^2 \triangleq \frac{1}{p}\sum_{i=1}^{p}\sigma_i^2$ and $w_i \triangleq \sigma_i^2 - \sigma^2$. Now the model parameters vector of (30) can be expressed via:

$$\boldsymbol{\Theta}(k) = (\lambda_1, \ldots, \lambda_k, \mathbf{V}_1, \ldots, \mathbf{V}_k, \sigma, w_1, \ldots, w_p, \beta_k). \tag{56}$$

This modification requires changing steps 3 and 4 in [34] (p. 3547) as follows:

In Step 3: Adding the computation of β_k as follows:

$$\beta_k = \text{mean}\left(\text{offdiag}\left(\hat{\mathbf{R}} - \mathbf{A}_k\mathbf{R}_{s,k}(\mathbf{A}_k)^H - (\sigma_{n,k})^2\mathbf{I}\right)\right), \tag{57}$$

where offdiag$(\mathbf{R})$ returns a second diagonal of the matrix $\mathbf{R}$.

In step 4: Changing the computation of $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{w}_k$ to $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{w}_k - \beta_k\mathbf{I}_{\text{off}}$, where $\mathbf{I}_{\text{off}}$ denotes a $p \times p$ matrix with ones on its second diagonals and zeros everywhere else.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful and constructive comments.

REFERENCES

[1] 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.*

[2] Y. P. Hong, C.T. Pan, "Rank-Revealing QR Factorizations and the Singular Value Decomposition", *Mathematics of Computation, vol. 58, No. 197 (Jan. 1992), pp. 213-232.*

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

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

[5] 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.*

[6] Q. Du, C. I. Chang, "Noise subspace projection approaches to determination of intrinsic dimensionality of hyperspectral imagery ", *Proc. Image and Signal Processing for Remote Sensing, SPIE Vol. 3871, pp. 34-44, December 1999.*

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

[8] 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.*

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

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

[11] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-33, pp. 387-392, Apr. 1985.*

[12] J. Rissanen "Modeling by shortest data description," *Automatica, vol. 14, no. 5, pp. 465471, 1978.*

[13] J. Rissanen "Estimation of structure by minimum description length," *Circuits, Syst. Signal Process., vol. 1, no. 4, pp. 395406, 1982.*

[14] P. Stoica and Y. Selen. "Model-order selection: a review of information criterion rules", *Signal Processing Magazine, IEEE, vol. 21, Issue 4, July 2004, pp. 36-47.*

[15] E. Moulines, P. Duhamel, J. F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing, vol. 43, pp. 516 525, Feb. 1995.*

[16] 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.*

[17] D. Slock, "Blind fractionally-spaced equalization, perfect reconstruction filterbanks, and multilinear prediction" *in Proc. ICASSP, Adelaide, Australia, Apr. 1994.*

[18] Y. Wu and K. W. Tam "On Determination of the Number of Signals in Spatially Correlated Noise," *IEEE Trans. Signal Processing, vol. 46, no. 11, pp. 3023 - 3029 November 1998.*

[19] K. M. Wong, Q. T. Zhang, J. P. Reilly, and P. Yip, , "A new criterion for the determination of the number of signals in high-resolution array processing", *Advanced algorithms and architectures for signal processing III; Proceedings of the Meeting, San Diego, CA, Aug. 15-17, 1988, pp. 352 - 357*

[20] A. P. Liavas, P. A. Regalia and J. P. Delmas, "Blind Channel Approximation: Effective Channel Order Determination," *IEEE Trans. Signal Processing, vol. 47, no. 12, December 1999.*

[21] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis,* Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[22] S. A. Kassam, H. V. Poor "Robust Techniques for Signal Processing: A Survey," *Proceedings of the IEEE Vol.73, Issue 3, March 1985 Page(s):433 - 481.*

[23] N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *Applied Statistics*, 29(3):231 2137, January 1980.

[24] 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.

[25] M. Leadbetter, *Extremes and related properties of random sequences and processes*, Springer Series in Statistics, 1982.

[26] S. I. Resnick, *Extreme Values, Regular Variation, and Point Process*, Springer Verlag, New York., 1987.

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

[28] E. J. Gumbel, *Statistics of Extremes.*, Columbia University Press., 1958.

[29] A. Hyvrinen. "Survey on Independent Component Analysis" *Neural Computing Surveys 2, pp. 94 - 128, 1999.*

[30] M. Abramowitz, I.A. Stegun "Handbook of Mathematical Functions" *Dover. Section 26.4.25, 1972.*

[31] S. Boyd, L. Vandenberghe "Convex Optimization" *Cambridge University Press, March 2004.*

[32] N. Johson, S. Kotz "Distributions in Statstics: Continuous Univariate Distributions-2" *John Wiley and Sons, 1970, pp. 130-148.*

[33] I. K. Fodor "A survey of dimension reduction techniques" *LLNL technical report, June 2002.*

[34] E. Fishler and H. V. Poor "Estimation of the Number of Sources in Unbalanced Arrays via Information Theoretic Criteria" *IEEE Trans. on signal processing, vol. 53, no. 9, Sept. 2005, pp. 3543-3553.*