CS decomposition based Bayesian subspace estimation

Olivier Besson*, Nicolas Dobigeon† and Jean-Yves Tourneret

February 15, 2012

Abstract

In numerous applications, it is required to estimate the principal subspace of the data, possibly from a very limited number of samples. Additionally, it often occurs that some rough knowledge about this subspace is available and could be used to improve subspace estimation accuracy in this case. This is the problem we address herein and, in order to solve it, a Bayesian approach is proposed. The main idea consists of using the CS decomposition of the semi-orthogonal matrix $\mathbf{H}$ whose columns span the subspace of interest. This parametrization is intuitively appealing and allows for non informative prior distributions of the matrices involved in the CS decomposition and very mild assumptions about the angles between the actual subspace and the prior subspace. The posterior distributions are derived and a Gibbs sampling scheme is presented to obtain the minimum mean-square distance estimator of the subspace of interest. Numerical simulations and an application to real hyperspectral data assess the validity and the performances of the estimator.

*O. Besson is with the University of Toulouse, ISAE, Department Electronics Optronics Signal, Toulouse (e-mail: olivier.besson@isae.fr). The work of O. Besson was partly supported by DGA-MRIS under grant no. 2009.60.033.00.470.75.01.

†N. Dobigeon and J.-Y. Tourneret are with the University of Toulouse, IRIT/ENSEEIHT, Signal and Communications Group, Toulouse, France (e-mail: nicolas.dobigeon@enseeiht.fr, jean-yves.tourneret@enseeiht.fr).

1 Problem statement

The ubiquitous linear model [1, 2], where the N-dimensional received signal can be written as a linear combination of p basis functions embedded in noise, has received a huge amount of attention due to its simplicity and relevance in a large number of applications. These applications include hyperspectral imagery which will be further investigated later in this paper. Under this framework, the $N \times K$ observation matrix $\mathbf{X}$, where N is the dimension of the observation space and K denotes the number of measurements, can be decomposed as

$$\mathbf{X} = \mathbf{H}\mathbf{\Psi} + \mathbf{N} \tag{1}$$

where $\mathbf{H}$ is an $N \times p$ matrix whose columns span the p-dimensional subspace of interest, $\mathbf{\Psi}$ is a $p \times K$ matrix whose columns correspond to the coordinates of the signal in the range space $\mathcal{R}(\mathbf{H})$ of $\mathbf{H}$, and $\mathbf{N}$ denotes the additive noise. In this paper, contrary to plenty of source separation techniques such as non-negative matrix factorization or independent component analysis, we are not interested in factorizing $\mathbf{X}$ into a product of unknown matrices $\mathbf{H}\mathbf{\Psi}$. Conversely, the problem addressed in this work consists of estimating the p-dimensional subspace of interest $\mathcal{R}(\mathbf{H})$, which is spanned by the columns of $\mathbf{H}$. As a consequence, without loss of generality, we assume in the sequel that the columns of $\mathbf{H}$ are orthonormal, i.e., $\mathbf{H}^T\mathbf{H} = \mathbf{I}$. When the columns of $\mathbf{N}$ are independent and Gaussian distributed with zero mean and covariance matrix $\sigma^2\mathbf{I}_N$, the maximum likelihood (ML) estimate of $\mathcal{R}(\mathbf{H})$ is obtained from the p most significant left singular vectors of $\mathbf{X}$ [1]. Therefore, the singular value decomposition (SVD) plays a central role in subspace estimation (in the frequentist framework) as it naturally reveals the low-rank structure of the signal. The SVD turns out to provide very accurate estimates of $\mathcal{R}(\mathbf{H})$ in most cases [3–5]. However, two situations of practical interest may undermine it. The first situation corresponds to the low sample regime, a case of most interest to us as will be evidenced in the hyperspectral application of Section 4. When K is small the SVD may not produce reliable estimates: this phenomenon is especially pronounced in large dimensional problems where K might be much lower than N. In this case, the sample covariance matrix is rank-deficient and its principal subspace is poorly estimated. In order to restore a better conditioned and more accurate covariance matrix estimate, numerous techniques have been proposed including shrinkage [6], dimensionality reduction using random unitary matrices [7], constrained maximum likelihood estimation (see e.g., [8] where the matrix of eigenvectors is constrained to be a product of Givens rotations) or eigenspace estimation using random matrix theory [9]. In the present paper, we even consider the situation where the number of snapshots K is less than the subspace dimension p. In this case, the SVD by itself is not sufficient as $\mathbf{X}$ is at most of rank $K < p$ and therefore it becomes impossible to recover $\mathcal{R}(\mathbf{H})$ without any further information. Another problem arises when the signal to noise ratio (SNR) is low and hence the separation between signal singular values and noise singular values is not clear. This may result in leakage of the signal subspace into the noise subspace, or even to a subspace swap, which leads to very inaccurate subspace estimates. This phenomenon has been evidenced e.g., in [10, 11] and theoretical explanations, based on the theory of large dimensional random matrices [12] are now available to predict this behavior [13–15]. In fact, for the two cases mentioned previously, additional prior information may prove to be helpful, and this prior information is often available either through models, expertise or data (cf. the hyperspectral application studied later in this paper). A natural way to introduce such knowledge is to adhere to a Bayesian framework. This is the approach we advocate in the present paper where our main focus is on *knowledge-aided subspace estimation in the low sample support or low SNR regime*.

More precisely, we assume that $\mathbf{H}$ is assigned some prior distribution $\pi(\mathbf{H})$, and our goal is to estimate $\mathbf{H}$ from the posterior distribution $p(\mathbf{H}|\mathbf{X})$. Similarly to [16, 17] we consider minimum mean square distance (MMSD) estimators of $\mathbf{H}$, i.e., we look for estimates $\hat{\mathbf{H}}$ of $\mathbf{H}$ that minimize the average squared Frobenius norm of the difference between the projection matrices, viz $\mathsf{E}\left\{\left\|\hat{\mathbf{H}}\hat{\mathbf{H}}^T - \mathbf{H}\mathbf{H}^T\right\|_F^2\right\}$.

The rationale behind this approach is that the usual mean-square metric $\mathsf{E}\left\{\left\|\hat{\mathbf{H}}-\mathbf{H}\right\|_F^2\right\}$ is not the natural metric on the Stiefel manifold [18, 19] while the distance between projection matrices is meaningful[1]. Using the latter distance, the MMSD estimator was shown to be given by [16, 17]

$$\begin{aligned}\hat{\mathbf{H}}_{\text{mmsd}} &= \arg\max_{\hat{\mathbf{H}}} \mathsf{E}\left\{\text{Tr}\left\{\hat{\mathbf{H}}^T\mathbf{H}\mathbf{H}^T\hat{\mathbf{H}}\right\}\right\} \\ &= \arg\max_{\hat{\mathbf{H}}} \int \text{Tr}\left\{\hat{\mathbf{H}}^T\mathbf{H}\mathbf{H}^T\hat{\mathbf{H}}\right\} p\left(\mathbf{H}|\mathbf{X}\right) d\mathbf{H} \\ &= \arg\max_{\hat{\mathbf{H}}} \text{Tr}\left\{\hat{\mathbf{H}}^T\left[\int \mathbf{H}\mathbf{H}^T p\left(\mathbf{H}|\mathbf{X}\right) d\mathbf{H}\right]\hat{\mathbf{H}}\right\} \\ &= \mathcal{P}_p\left\{\int \mathbf{H}\mathbf{H}^T p\left(\mathbf{H}|\mathbf{X}\right) d\mathbf{H}\right\} \end{aligned} \tag{2}$$

where $\mathcal{P}_p\{\cdot\}$ stands for the p principal eigenvectors of the matrix between braces. The MMSD estimator thus amounts to finding the principal subspace of the posterior mean of the projection matrix $\mathbf{P}=\mathbf{H}\mathbf{H}^T$ on $\mathcal{R}(\mathbf{H})$. Note that this approach is general and independent of the conditional and prior distributions: depending on the latter, it may or may not be an easy task to obtain the MMSD estimator. In the sequel, we state our assumptions regarding $\mathbf{H}$ and derive its corresponding MMSD estimator. The latter will then be tested on real hyperspectral data in Section 4.

2 Data model and subspace estimation

Let us consider the linear model (1) and let us assume that $\mathbf{N}$ is Gaussian distributed with independent columns so that the probability density function of $\mathbf{X}$, conditioned on $\mathbf{H}$ and $\mathbf{\Psi}$, is given by

$$p\left(\mathbf{X}|\mathbf{H},\mathbf{\Psi}\right) \propto \text{etr}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{X}-\mathbf{H}\mathbf{\Psi}\right)^T\left(\mathbf{X}-\mathbf{H}\mathbf{\Psi}\right)\right\} \tag{3}$$

where $\text{etr}\{\cdot\}$ stands for the exponential of the trace of the matrix between braces and $\propto$ means proportional to. Since the thermal noise level can usually be estimated with high accuracy, we assume here that σ^2 is known[2]. Since no knowledge about $\mathbf{\Psi}$ is generally available, we treat it as a random matrix with uniform prior distribution, i.e., $\pi\left(\mathbf{\Psi}\right) \propto 1$, so that the distribution of $\mathbf{X}$, conditioned on

[1] The true (square) distance between the subspaces is given by $d^2\left(\hat{\mathbf{H}},\mathbf{H}\right) = \sum_{k=1}^p \theta_k^2$ where θ_k for $k=1,\ldots,p$ stand for the principal angles between $\mathcal{R}\left(\hat{\mathbf{H}}\right)$ and $\mathcal{R}\left(\mathbf{H}\right)$. The distance we use herein, i.e., $\left\|\hat{\mathbf{H}}\hat{\mathbf{H}}^T-\mathbf{H}\mathbf{H}^T\right\|_F^2 = 2p - 2\text{Tr}\left\{\hat{\mathbf{H}}^T\mathbf{H}\mathbf{H}^T\hat{\mathbf{H}}\right\} = 2\sum_{k=1}^p \sin^2\theta_k$, is thus different from $d^2\left(\hat{\mathbf{H}},\mathbf{H}\right)$. However, the two distances are close for small values of θ_k and the distance between projection matrices is widely accepted. Moreover, using the distance between projection matrices allows one to obtain a closed-form expression for the MMSD estimator, see (2). Minimization of $\mathsf{E}\left\{d^2\left(\hat{\mathbf{H}},\mathbf{H}\right)\right\}$ would not yield such closed-form expression since $\sum_{k=1}^p \theta_k^2$ cannot be expressed simply as a function of $\hat{\mathbf{H}}$ and $\mathbf{H}$.

[2] The case of unknown σ^2 can be considered by assigning a prior distribution (typically a conjugate prior, in our case an inverse gamma distribution) to σ^2 and modifying accordingly the posterior distributions to be derived next.

$\mathbf{H}$ only, is obtained as

$$\begin{aligned} p\left(\mathbf{X}|\mathbf{H}\right) &= \int p\left(\mathbf{X}|\mathbf{H},\mathbf{\Psi}\right)\pi\left(\mathbf{\Psi}\right)d\mathbf{\Psi} \\ &\propto \int \mathrm{etr}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{X}-\mathbf{H}\mathbf{\Psi}\right)^T\left(\mathbf{X}-\mathbf{H}\mathbf{\Psi}\right)\right\}d\mathbf{\Psi} \\ &\propto \mathrm{etr}\left\{-\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{H}\mathbf{H}^T\mathbf{X}\right\} \\ &\times \int \mathrm{etr}\left\{-\frac{1}{2\sigma^2}\left(\mathbf{\Psi}-\mathbf{H}^T\mathbf{X}\right)^T\left(\mathbf{\Psi}-\mathbf{H}^T\mathbf{X}\right)\right\}d\mathbf{\Psi} \\ &\propto \mathrm{etr}\left\{-\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X}+\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{H}\mathbf{H}^T\mathbf{X}\right\} \end{aligned} \tag{4}$$

where, to obtain the last line, we have used the fact that the integral in the fourth line of (4) is that of a multivariate Gaussian distribution with mean $\mathbf{H}^T\mathbf{X}$ and covariance matrix $\sigma^2\mathbf{I}$, and hence is proportioanl to σ^{pK}. Note that $p\left(\mathbf{X}|\mathbf{H}\right)$ depends on $\mathbf{H}$ only through the projection matrix $\mathbf{P}=\mathbf{H}\mathbf{H}^T$.

Let us turn now to the hypotheses regarding $\mathbf{H}$. We assume that we have some a priori knowledge about the subspace spanned by the columns of $\mathbf{H}$: this knowledge can come from some available models or can be deduced from the data itself, as in the hyperspectral imagery application. More precisely, we assume that the range space $\mathcal{R}\left(\mathbf{H}\right)$ of $\mathbf{H}$ is close to the range space of some semi-orthogonal matrix $\bar{\mathbf{H}}$ and, without loss of generality, we will assume that $\bar{\mathbf{H}}=\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}^T$ through the paper[3].

In [17], we tackled the problem by assigning the matrix $\mathbf{H}$ either a Bingham –$\pi_{\mathrm{B}}\left(\mathbf{H}\right)\propto\mathrm{etr}\left\{\kappa\mathbf{H}^T\bar{\mathbf{H}}\bar{\mathbf{H}}^T\mathbf{H}\right\}$– or a von Mises Fisher (vMF) distribution – $\pi_{\mathrm{vMF}}\left(\mathbf{H}\right)\propto\mathrm{etr}\left\{\kappa\mathbf{H}^T\bar{\mathbf{H}}\right\}$. The Bingham and vMF are the most widely used distributions on the Stiefel manifold and they have proved to be relevant in a number of applications, including meteorology, biology, image or shape analysis [20]. Moreover, there exists computationally efficient simulation tools to sample from these distributions, which makes them a sensible choice. However, they suffer from two drawbacks. First, from a user point of view, it is not obvious to set a value for the concentration parameter κ since the latter is not an intuitively appealing parameter, in contrast to the angles between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ which are more directly meaningful. Moreover, the Bingham and vMF distributions hold for the whole matrix $\mathbf{H}$: the choice of a distribution and a value for κ will consequently induce a distribution for the angles, but this relation is not revealed in a straightforward and intelligible manner. In the present paper, we attempt to remedy these shortcomings with a view to obtain a parametrization of the statistical model that directly involves the most meaningful parameters, namely the angles θ_k, $k=1,\cdots,p$ between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$. Indeed, these angles are instrumental as the distance between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ is directly connected to them. Furthermore, we look for a less constrained model which relies on mild assumptions, and the latter would only concern the angles θ_k.

The model proposed herein is based on the CS decomposition of $\mathbf{H}$, which writes [19]

$$\mathbf{H}=\begin{bmatrix}\mathbf{U}_1\mathbf{C}\\ \mathbf{U}_2\mathbf{S}\end{bmatrix}\mathbf{V}^T \tag{5}$$

where $\mathbf{U}_1$ and $\mathbf{V}$ are $p\times p$ orthogonal matrices, $\mathbf{U}_2$ is an $(N-p)\times p$ semi-orthogonal matrix ($\mathbf{U}_2^T\mathbf{U}_2=\mathbf{I}_p$), $\mathbf{C}=\mathrm{diag}\left(\cos\theta_1,\cdots,\cos\theta_p\right)$ and $\mathbf{S}=\mathrm{diag}\left(\sin\theta_1,\cdots,\sin\theta_p\right)$. The angles θ_k correspond to the principal angles between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ while the columns of $\begin{bmatrix}\mathbf{U}_1\\ \mathbf{0}\end{bmatrix}$ and $\mathbf{H}\mathbf{V}$ are the associated principal vectors. As requested, this representation has the nice property that the angles between

[3]In the case where $\mathbf{H}$ is close to an arbitrary semi-orthogonal matrix $\bar{\mathbf{H}}$, the measurements in (1) can be pre-multiplied by the unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}\bar{\mathbf{H}}=\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}^T$. Note that pre-multiplication by the unitary matrix $\mathbf{Q}$ does not modify the angles between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ nor the distribution $p\left(\mathbf{X}|\mathbf{H},\mathbf{\Psi}\right)$ in (3).

$\mathcal{R}(\mathbf{H})$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ are directly revealed, and do not depend on the matrices $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{V}$, which can be arbitrary. We now assign prior distributions to the model variables. First observe that the likelihood function in (4) depends on $\mathbf{H}$ only through the projection matrix $\mathbf{P} = \mathbf{H}\mathbf{H}^T$ and the latter, under the CS decomposition (5), is independent of $\mathbf{V}$. Therefore, we need to set prior distributions for $\mathbf{U}_1$, $\mathbf{U}_2$ and $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \cdots & \theta_p\end{bmatrix}^T$ only. As for $\mathbf{U}_1$ and $\mathbf{U}_2$ we assume that they have uniform prior distributions on the orthogonal group $\mathcal{O}(p)$ and the Stiefel manifold $\mathcal{S}_{p,N-p}$, i.e., the set of $(N-p)\times p$ matrices $\mathbf{U}_2$ such that $\mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}_p$. As for $\boldsymbol{\theta}$, we assume that θ_k are independent and identically distributed (i.i.d.) random variables, with uniform distribution on $[0, \theta_{\max}]$, i.e., $\theta_k \sim \mathrm{U}([0, \theta_{\max}])$. Observe that, as stated in our objectives, the statistical model involves rather mild assumptions. Moreover, it directly involves the angles θ_k, which makes sense intuitively. Finally, the only parameter the user has to set is $\theta_{\max}$, which seems easier to set than a value for κ. Indeed $\theta_{\max}$ rules the maximum angle between $\mathcal{R}(\mathbf{H})$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$: therefore, the smaller $\theta_{\max}$, the closer these subspaces a priori. In contrast, when $\theta_{\max}$ increases, the two subspaces can be quite far apart. Consequently, for small $\theta_{\max}$ we can expect the MMSD estimator to strongly rely on $\bar{\mathbf{H}}$, while for large $\theta_{\max}$ the data $\mathbf{X}$ is likely to prevail.

Since the likelihood and the prior distributions have been set, we now consider the posterior distributions of $\mathbf{U}_1$, $\mathbf{U}_2$ and $\boldsymbol{\theta}$. As a preliminary step, note that

$$\mathbf{P} = \mathbf{H}\mathbf{H}^T = \begin{bmatrix}\mathbf{U}_1\mathbf{C}^2\mathbf{U}_1^T & \mathbf{U}_1\mathbf{C}\mathbf{S}\mathbf{U}_2^T \\ \mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T & \mathbf{U}_2\mathbf{S}^2\mathbf{U}_2^T\end{bmatrix} \tag{6}$$

so that, with the partitioning $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1^T & \mathbf{X}_2^T\end{bmatrix}^T$, we have

$$\begin{aligned}\mathrm{Tr}\left\{\mathbf{X}^T\mathbf{P}\mathbf{X}\right\} &= \mathrm{Tr}\left\{\begin{bmatrix}\mathbf{X}_1^T & \mathbf{X}_2^T\end{bmatrix}\begin{bmatrix}\mathbf{U}_1\mathbf{C}^2\mathbf{U}_1^T & \mathbf{U}_1\mathbf{C}\mathbf{S}\mathbf{U}_2^T \\ \mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T & \mathbf{U}_2\mathbf{S}^2\mathbf{U}_2^T\end{bmatrix}\begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix}\right\} \\ &= \mathrm{Tr}\left\{\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1 + 2\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T\mathbf{X}_1 + \mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right\}. \end{aligned} \tag{7}$$

Assuming a priori independence between $\mathbf{U}_1$, $\mathbf{U}_2$ and $\boldsymbol{\theta}$, it follows from (4) that the joint posterior distribution of $\mathbf{U}_1$, $\mathbf{U}_2$ and $\boldsymbol{\theta}$ is given by

$$\begin{aligned}p(\mathbf{U}_1, \mathbf{U}_2, \boldsymbol{\theta}|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{H})\,\pi(\mathbf{U}_1)\,\pi(\mathbf{U}_2)\,\pi(\boldsymbol{\theta}) \\ &\propto \mathrm{etr}\left\{\frac{1}{2\sigma^2}\left[\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1 + 2\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T\mathbf{X}_1 + \mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right]\right\}\pi(\mathbf{U}_1)\pi(\mathbf{U}_2)\pi(\boldsymbol{\theta}).\end{aligned} \tag{8}$$

In order to obtain the MMSD estimator, we suggest, as in [17], to use a Gibbs sampler which enables one to iteratively draw samples from the posterior distribution of each variable, conditioned on all other variables [21, 22]. In order to obtain the conditional posterior distribution of $\mathbf{U}_1$ only, we start with (8) and note that

$$\begin{aligned}p(\mathbf{U}_1, \mathbf{U}_2, \boldsymbol{\theta}|\mathbf{X}) &\propto \mathrm{etr}\left\{\frac{1}{2\sigma^2}\left[\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1 + 2\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T\mathbf{X}_1 + \mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right]\right\}\pi(\mathbf{U}_1)\pi(\mathbf{U}_2)\pi(\boldsymbol{\theta}) \\ &\propto \mathrm{etr}\left\{\frac{1}{2\sigma^2}\left[\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1 + 2\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T\mathbf{X}_1\right]\right\}\pi(\mathbf{U}_1) \\ &\times \mathrm{etr}\left\{\frac{1}{2\sigma^2}\left[\mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right]\right\}\pi(\mathbf{U}_2)\pi(\boldsymbol{\theta}) \\ &= f(\mathbf{U}_1, \mathbf{U}_2, \boldsymbol{\theta}) \times g(\mathbf{U}_2, \boldsymbol{\theta}).\end{aligned}$$

Therefore,

$$
\begin{aligned}
p\left(\mathbf{U}_1|\mathbf{U}_2,\boldsymbol{\theta},\mathbf{X}\right) &= \frac{p\left(\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta}|\mathbf{X}\right)}{\int p\left(\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta}|\mathbf{X}\right)d\mathbf{U}_1} \\
&= \frac{f\left(\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta}\right)g(\mathbf{U}_2,\boldsymbol{\theta})}{\int f((\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta})\,g(\mathbf{U}_2,\boldsymbol{\theta})d\mathbf{U}_1} \\
&= \frac{f\left(\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta}\right)}{\int f((\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta})\,d\mathbf{U}_1} \\
&\propto f\left(\mathbf{U}_1,\mathbf{U}_2,\boldsymbol{\theta}\right) \\
&\propto \operatorname{etr}\left\{\frac{1}{2\sigma^2}\left[2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}+\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1\right]\right\}\mathbb{I}_{\mathcal{O}(p)}\left(\mathbf{U}_1\right) \qquad (9)
\end{aligned}
$$

where $\mathbb{I}_{\mathcal{O}(p)}\left(\mathbf{U}_1\right)$ is the indicator function defined on $\mathcal{O}(p)$ (i.e., $\mathbb{I}_{\mathcal{O}(p)}\left(\mathbf{U}_1\right)=1$ if $\mathbf{U}_1\in\mathcal{O}(p)$ and 0 otherwise). The distribution in (9) is recognized as a Bingham-von-Mises-Fisher (BMF) distribution with parameter matrices $\mathbf{X}_1\mathbf{X}_1^T$, $\frac{1}{2\sigma^2}\mathbf{C}^2$ and $\frac{1}{\sigma^2}\mathbf{X}_1\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}$ respectively[4]. An efficient sampling scheme to generate random matrices drawn from a BMF $(\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3)$ distribution on the Stiefel manifold was proposed in [23]. In our case, $\mathbf{U}_1\in\mathcal{O}(p)$ and, as mentioned in [23], the sampling scheme on the Stiefel manifold cannot be used directly and needs to be modified. In Appendix A, following the lines of [23], we give some details about the sampling scheme for the matrix BMF distribution on the orthogonal group $\mathcal{O}(p)$. Similarly, In order to obtain the conditional posterior distribution of $\mathbf{U}_2$ only, we start with (8) and keep only the terms which depend on $\mathbf{U}_2$ since the other terms will appear as constants and can be absorbed in the normalization constant. Doing so, we deduce that

$$
p\left(\mathbf{U}_2|\mathbf{U}_1,\boldsymbol{\theta},\mathbf{X}\right)\propto\operatorname{etr}\left\{\frac{1}{2\sigma^2}\left[2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_1^T\mathbf{U}_1\mathbf{C}\mathbf{S}+\mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right]\right\}\mathbb{I}_{\mathcal{S}_{p,N-p}}\left(\mathbf{U}_2\right) \qquad (10)
$$

and hence

$$
\mathbf{U}_2|\mathbf{U}_1,\boldsymbol{\theta},\mathbf{X}\sim\text{BMF}\left(\mathbf{X}_2\mathbf{X}_2^T,\frac{1}{2\sigma^2}\mathbf{S}^2,\frac{1}{\sigma^2}\mathbf{X}_2\mathbf{X}_1^T\mathbf{U}_1\mathbf{C}\mathbf{S}\right). \qquad (11)
$$

Since $\mathbf{U}_2\in\mathcal{S}_{p,N-p}$, the sampling scheme of Hoff [23] can be used to draw matrices from the distribution in (11).

Let us now examine the posterior distribution of $\boldsymbol{\theta}$

$$
\begin{aligned}
p\left(\boldsymbol{\theta}|\mathbf{U}_1,\mathbf{U}_2,\mathbf{X}\right) &\propto \operatorname{etr}\left\{\frac{1}{2\sigma^2}\left[\mathbf{C}^2\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1+2\mathbf{X}_2^T\mathbf{U}_2\mathbf{S}\mathbf{C}\mathbf{U}_1^T\mathbf{X}_1+\mathbf{S}^2\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2\right]\right\}\pi(\boldsymbol{\theta}) \\
&\propto \prod_{k=1}^{p}\exp\left\{\alpha_k\cos^2\theta_k+2\beta_k\cos\theta_k\sin\theta_k+\gamma_k\sin^2\theta_k\right\}\mathbb{I}_{[0,\theta_{\max}]}(\theta_k) \qquad (12)
\end{aligned}
$$

where α_k, β_k, γ_k are the k-th diagonal entries of $\frac{1}{2\sigma^2}\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_1^T\mathbf{U}_1$, $\frac{1}{2\sigma^2}\mathbf{U}_1^T\mathbf{X}_1\mathbf{X}_2^T\mathbf{U}_2$ and $\frac{1}{2\sigma^2}\mathbf{U}_2^T\mathbf{X}_2\mathbf{X}_2^T\mathbf{U}_2$, respectively. The first thing to be noted is that the variables θ_k, conditioned on $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{X}$, are independent and hence one needs to generate p independent random variables. Unfortunately, the distribution in (12) does not belong to any known class of distributions and, therefore, generating random variables drawn from $p\left(\boldsymbol{\theta}|\mathbf{U}_1,\mathbf{U}_2,\mathbf{X}\right)$ appears problematic. In order to overcome this problem, we propose to resort to a Metropolis-Hastings (MH) move [21, 22]. The basic idea is to generate a random variable drawn from a proposal distribution and to accept it with a certain probability, the latter being equal to one if the candidate contributes to increase the target posterior distribution. Of course, the closer the proposal and target distributions, the higher the acceptance rate and hence the faster the convergence of the Markov chain. In order to obtain a proposal distribution in our case, we

[4]The matrix $\mathbf{H}\in\mathcal{S}_{p,q}$ is said to have a BMF $(\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3)$ distribution -where $\mathbf{A}_1$ is an $q\times q$ symmetric matrix, $\mathbf{A}_2$ is a $p\times p$ diagonal matrix and $\mathbf{A}_3$ is an $q\times p$ matrix- if $p(\mathbf{H})\propto\operatorname{etr}\left\{\mathbf{H}^T\mathbf{A}_3+\mathbf{A}_2\mathbf{H}^T\mathbf{A}_1\mathbf{H}\right\}$.

make the change of variable $x_k = \sin^2\theta_k$ in (12), and come up with the equivalent problem of finding a proposal distribution for the conditional distribution of x_k, which is is given by

$$p\left(x_k|\mathbf{U}_1,\mathbf{U}_2,\mathbf{X}\right) \propto x_k^{-1/2}(1-x_k)^{-1/2}\exp\left\{-(\alpha_k-\gamma_k)x_k + 2\beta_k x_k^{1/2}(1-x_k)^{1/2}\right\}\mathbb{I}_{[0,x_{\max}]}(x_k) \quad (13)$$

where $x_{\max} = \sin^2\theta_{\max}$. Forgetting the exponential term in (13), this distribution is similar to that of a scaled beta distribution. Therefore, we choose a scaled beta distribution $q(x_k) \propto \left(\frac{x_k}{x_{\max}}\right)^{a_k-1}(1-\frac{x_k}{x_{\max}})^{b_k-1}$ as a proposal distribution in a Metropolis Hastings scheme. Through preliminary investigation, we ended up with the choice $a_k = 0.5 + 0.25\max(0,\beta_k) - 0.25\min(0,\alpha_k-\gamma_k)$ and $b_k = 0.5 + 0.25\max(0,\beta_k) + 0.25\max(0,\alpha_k-\gamma_k)$ which turns out to provide a good approximation to (13) for low to moderate SNR. The resulting Gibbs sampling scheme is summarized in Algorithm 1.

Algorithm 1 Gibbs sampler for estimation of $\mathbf{H}$ using the CS decomposition.

Input: initial values $\mathbf{U}_1^{(0)}$, $\mathbf{U}_2^{(0)}$, $\boldsymbol{\theta}^{(0)}$

1: **for** $n = 1,\cdots,N_{\text{bi}}+N_r$ **do**
2: sample $\mathbf{U}_1^{(n)}$ from BMF $\left(\mathbf{X}_1\mathbf{X}_1^T, \frac{1}{2\sigma^2}(\mathbf{C}^{(n-1)})^2, \frac{1}{\sigma^2}\mathbf{X}_1\mathbf{X}_2^T\mathbf{U}_2^{(n-1)}\mathbf{S}^{(n-1)}\mathbf{C}^{(n-1)}\right)$ in (9) where $\mathbf{C}^{(n-1)} = \text{diag}\left(\cos\boldsymbol{\theta}^{(n-1)}\right)$ and $\mathbf{S}^{(n-1)} = \text{diag}\left(\sin\boldsymbol{\theta}^{(n-1)}\right)$.
3: sample $\mathbf{U}_2^{(n)}$ from BMF $\left(\mathbf{X}_2\mathbf{X}_2^T, \frac{1}{2\sigma^2}(\mathbf{S}^{(n-1)})^2, \frac{1}{\sigma^2}\mathbf{X}_2\mathbf{X}_1^T\mathbf{U}_1^{(n)}\mathbf{C}^{(n-1)}\mathbf{S}^{(n-1)}\right)$ in (10).
4: **Metropolis-Hastings** to sample $\boldsymbol{\theta}^{(n)}$:
5: **for** $k = 1,\cdots,p$ **do**
6: draw an initial candidate $x_k^{c(0)}$ from $x_{\max}\times\text{Beta}\left(a_k,b_k\right)$ and set $x_k^{(n)} = x_k^{c(0)}$.
7: **for** $\ell = 1,\cdots,q$ **do**
8: draw a candidate $x_k^{c(\ell)}$ from $x_{\max}\times\text{Beta}\left(a_k,b_k\right)$.
9: accept $x_k^{c(\ell)}$ as $x_k^{(n)}$ with probability $\min\left(1, \frac{p\left(x_k^{c(\ell)}|\mathbf{U}_1^{(n)},\mathbf{U}_2^{(n)},\mathbf{X}\right)}{p\left(x_k^{c(\ell-1)}|\mathbf{U}_1^{(n)},\mathbf{U}_2^{(n)},\mathbf{X}\right)}\frac{q(x_k^{c(\ell-1)})}{q(x_k^{c(\ell)})}\right)$ where $p\left(x_k|\mathbf{U}_1^{(n)},\mathbf{U}_2^{(n)},\mathbf{X}\right)$ is given in (13).
10: **end for**
11: $\theta_k^{(n)} = \arcsin\sqrt{x_k^{(n)}}$.
12: **end for**
13: **end for**

Output: sequence of random matrices $\mathbf{H}^{(n)} = \begin{bmatrix}\mathbf{U}_1^{(n)}\mathbf{C}^{(n)}\\ \mathbf{U}_2^{(n)}\mathbf{S}^{(n)}\end{bmatrix}$.

Once the matrices $\mathbf{H}^{(n)}$ have been generated, the MMSD estimator which theoretically entails computing $\mathcal{P}_p\left\{\int \mathbf{H}\mathbf{H}^T p\left(\mathbf{H}|\mathbf{X}\right)d\mathbf{H}\right\}$ can be approximated by

$$\hat{\mathbf{H}}_{\text{mmsd}} = \mathcal{P}_p\left\{\frac{1}{N_r}\sum_{n=N_{\text{bi}}+1}^{N_{\text{bi}}+N_r}\mathbf{H}^{(n)}\left(\mathbf{H}^{(n)}\right)^T\right\}. \quad (14)$$

Remark 1. Similarly, a maximum a posteriori (MAP) approach can be advocated where the MAP estimator is obtained as

$$\begin{aligned}\hat{\mathbf{H}}_{\text{map}} &= \arg\max_{\mathbf{H}^{(n)}} p\left(\mathbf{U}_1^{(n)},\mathbf{U}_2^{(n)},\boldsymbol{\theta}^{(n)}|\mathbf{X}\right)\\ &= \arg\max_{\mathbf{H}^{(n)}}\text{Tr}\left\{\mathbf{X}^T\mathbf{H}^{(n)}\left(\mathbf{H}^{(n)}\right)^T\mathbf{X}\right\}. \quad (15)\end{aligned}$$

Note that $\mathrm{Tr}\left\{\mathbf{X}^T\mathbf{H}\mathbf{H}^T\mathbf{X}\right\}$ is maximized when $\mathbf{H}$ is the matrix of the p most significant left singular vectors of $\mathbf{X}$ and, hence, the MAP approach is in some way linked to the SVD-based approach. Observe also that it does not make much sense to consider here a minimum mean-square error (MMSE) estimator. Indeed the latter entails computing $\int \mathbf{H}\, p\left(\mathbf{H}|\mathbf{X}\right) d\mathbf{H}$ which could be approximated by the arithmetic mean of the set of matrices $\mathbf{H}^{(n)}$. However, the range space of $\mathbf{H}$ is given up to right multiplication by an orthogonal matrix. Therefore, $\mathcal{R}\left(\mathbf{H}^{(n)}\right)$ could be close to $\mathcal{R}\left(\mathbf{H}\right)$ without the actual matrices $\mathbf{H}^{(n)}$ and $\mathbf{H}$ being close. It results that the arithmetic mean of the matrices $\mathbf{H}^{(n)}$ could result in a poor subspace estimate despite the fact that, individually, the subspaces spanned by each matrix $\mathbf{H}^{(n)}$ might be accurate.

3 Simulations

In this section we use Monte-Carlo simulations to assess the performance of the estimator defined previously. Towards this end, two figures of merit will be used, namely the average fraction of energy (AFE) of $\hat{\mathbf{H}}$ in $\mathcal{R}\left(\mathbf{H}\right)$ and the mean square distance (MSD) between the subspace spanned by $\hat{\mathbf{H}}$ and the subspace spanned by $\mathbf{H}$, where $\hat{\mathbf{H}}$ stands for one of the estimates. They are respectively given by

$$\mathrm{AFE}\left(\hat{\mathbf{H}},\mathbf{H}\right) = \frac{1}{p}\mathsf{E}\left\{\mathrm{Tr}\left\{\hat{\mathbf{H}}^T\mathbf{H}\mathbf{H}^T\hat{\mathbf{H}}\right\}\right\} = \frac{1}{p}\mathsf{E}\left\{\sum_{k=1}^{p}\cos^2\theta_k\right\} \tag{16}$$

$$\mathrm{MSD}\left(\hat{\mathbf{H}},\mathbf{H}\right) = \mathsf{E}\left\{d^2\left(\hat{\mathbf{H}},\mathbf{H}\right)\right\} = \mathsf{E}\left\{\sum_{k=1}^{p}\theta_k^2\right\}. \tag{17}$$

where θ_k, $k = 1,\ldots,p$ stand for the principal angles between $\mathcal{R}\left(\hat{\mathbf{H}}\right)$ and $\mathcal{R}\left(\mathbf{H}\right)$. In all simulations $N = 20$, $p = 5$ and $\bar{\mathbf{H}} = \begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}^T$. The matrix $\mathbf{\Psi}$ is generated from a Gaussian distribution with zero-mean and covariance matrix $\mathbf{I}_p$ and the signal-to-noise ratio is defined as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathsf{E}\left\{\mathrm{Tr}\left\{\mathbf{\Psi}^T\mathbf{H}^T\mathbf{H}\mathbf{\Psi}\right\}\right\}}{\mathsf{E}\left\{\mathrm{Tr}\left\{\mathbf{N}^T\mathbf{N}\right\}\right\}}\right) = 10\log_{10}\left(\frac{p}{N\sigma^2}\right).$$

The angles between $\mathcal{R}\left(\mathbf{H}\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$ are fixed over all simulations and set to $\boldsymbol{\theta} = \begin{bmatrix}15^\circ & 25^\circ & 35^\circ & 45^\circ & 55^\circ\end{bmatrix}^T$ which results in AFE $\left(\bar{\mathbf{H}},\mathbf{H}\right) = 0.6509$ and MSD $\left(\bar{\mathbf{H}},\mathbf{H}\right) = 2.1704$. The matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are drawn randomly at each Monte-Carlo run. The number of burn-in iterations in the Gibbs sampler is set to $N_{\mathrm{bi}} = 10$ and $N_r = 1000$ samples are used to approximate the estimators following (14) and (15). The MMSD estimator (14) is compared with the MAP estimator (15), the usual SVD-based estimator and the sparse matrix transform (SMT) of [8]. In all figures, the solid black line represents AFE $\left(\bar{\mathbf{H}},\mathbf{H}\right)$ or MSD $\left(\bar{\mathbf{H}},\mathbf{H}\right)$, i.e., when $\hat{\mathbf{H}} = \bar{\mathbf{H}}$ and only the a priori knowledge is used, the data being discarded. In all simulations, the AFE and the MSD are evaluated from 500 Monte-Carlo trials.

We now successively investigate the influence of $\theta_{\max}$, K and SNR in Figures 1 to 9. The first observation to be made is that the MMSD estimator is rather insensitive to the choice of $\theta_{\max}$: this is an interesting feature, as it means that $\theta_{\max}$ need not be fixed with a high accuracy. Next, it can be observed that the MMSD estimator outperforms the usual SVD-based estimator and the SMT estimator, for small K and low SNR: under these conditions, it makes a sound use of the prior information and provides more accurate estimates. Note also that it performs better than the estimate $\hat{\mathbf{H}} = \bar{\mathbf{H}}$, and hence the prior by itself is not sufficient. Finally, we observe that SMT is approximately equivalent to the SVD estimator, and that the MAP estimator does not perform as well as the MMSD estimator.

4 Application to hyperspectral data

Hyperspectral imagery has recently emerged as a feasible and relevant technique for accurate observation of earth surfaces, either for agricultural or geographical purposes [24]. The diversity of the frequency response of each component of the illuminated scene makes it possible a fine understanding of the soil characteristics, and thus numerous studies have focused on information retrieval from multi-band data, see, e.g., [25–28]. A widely accepted model so far is that the image can be linearly decomposed as a combination of a few components, referred to as the endmembers [29]. One critical issue is thus to identify the subspace where the data lies together with the coordinates in this subspace, which provide the respective abundances, i.e., the proportion of the soil components. This can be achieved by well-known and computationally efficient techniques such as principal component analysis (PCA), a primordial asset to using the linear (or subspace) model. However, it may be argued that the linear model does not fully account for all physical phenomenon that give rise to the image, e.g., the possibly non-linear mixing of the components. In order to obtain a finer image analysis, non-linear models can be investigated [29] but generally at the price of a higher computational complexity. Furthermore, in most cases non-linear effects are not that important and an interesting alternative is to continue to resort to a linear model but at a local level (i.e., within a few pixels) rather than at the full image level. Doing so, one can characterize the data locally and track the evolution of the local subspaces in order to assess the degree of non-linearity.

The subspace estimation scheme developed above can fulfill this task and it is now tested against real hyperspectral data, acquired by the NASA spectro-imager AVIRIS over Moffett Field, CA, and over the Cuprite mining site, Nevada in 1997. More precisely, we consider a 50×50 sub-image in each case. As for the Moffett image, it contains partly a lake (upper part of the sub-image) and partly a coastal area (lower part of the sub-image) composed of soil and vegetation, see [30] for a more detailed description. The area of interest considered in the Cuprite sub-image is usually referred to as the "alunite hill" and has been for instance investigated in [31], see also [30] for more details. It is mainly composed of three geological materials: muscovite, alunite and kaolinite. The data is collected in $N = 183$ spectral bands for the Moffett image and $N = 189$ spectral bands for the Cuprite image, and we have thus a total of $L = 2500$ pixels. Under the linear mixing model and in the absence of noise, the data matrix $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_L\end{bmatrix}$ where $\mathbf{y}_\ell \in \mathbb{R}^N$ stands for the ℓ-th pixel, can be written as $\mathbf{Y} = \mathbf{MA}$ where $\mathbf{M} = \begin{bmatrix}\mathbf{m}_1 & \cdots \mathbf{m}_R\end{bmatrix}$ and $\mathbf{m}_r$, $r = 1, \cdots, R$ denotes the set of endmembers, i.e., the spectral signatures which best describe the soil components. In [30], it was shown that a value $R = 3$ was sufficient to obtain an accurate description of the data. The columns $\mathbf{a}_\ell = \begin{bmatrix}a_{\ell,1} & \cdots a_{\ell,R}\end{bmatrix}^T$ of the matrix $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \cdots \mathbf{a}_L\end{bmatrix}$ are the so-called abundances: they satisfy the positivity constraint $a_{\ell,r} \geq 0$ and the sum-to-one property, i.e., $\mathbf{a}_\ell^T \mathbf{1}_R = 1$ where $\mathbf{1}_R$ is the R-length vector whose elements are all equal to 1. In other words, the matrix $\mathbf{A}$ satisfies the constraint $\mathbf{A}^T \mathbf{1}_R = \mathbf{1}_L$. The pixels $\mathbf{y}_\ell$ thus belong to a simplex whose vertices are the R endmembers $\mathbf{m}_r$ [30]. Let $\mu = L^{-1} \sum_{\ell=1}^L \mathbf{y}_\ell$ denote the mean value of the pixels. Then, the centered data matrix $\mathbf{X} = \mathbf{Y} - \mu \mathbf{1}_L^T$ belongs to a p-dimensional subspace (with $p = R - 1$) which can be estimated by a number of techniques, including principal component analysis (PCA) [30].

Usually, PCA is performed on the whole image, which makes sense if the linear mixing model is in force for all pixels. Herein, we are interested in assessing the validity of this model at the pixel level. More precisely, the PCA on the whole image will provide us with the "average" subspace: the pixels $\mathbf{y}_\ell$ are then unitarily transformed ($\mathbf{y}_\ell \leftarrow \mathbf{Q}\mathbf{y}_\ell$) such that $\bar{\mathbf{H}} \leftarrow \mathbf{Q}\bar{\mathbf{H}} = \begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}^T$ and we are interested in the distance between $\bar{\mathbf{H}}$ and the subspace spanned by a pixel $\mathbf{y}_\ell$ and its few nearest pixels. If this distance is very small, then it is likely that the linear model described by $\bar{\mathbf{H}}$ is rather accurate. On the other hand, if the distance is not negligible, it may be that $\bar{\mathbf{H}}$ does not describe accurately the scene around pixel ℓ, or that some non-linear mixing effects might occur there. Therefore, subspace estimation at the pixel level together with distance to $\bar{\mathbf{H}}$ evaluation enables one to gain insight into the understanding of the mixing process. This is the approach we take here and our MMSD estimator

is used towards this end. To be more specific, for each pixel $\mathbf{y}_\ell$ we use the latter and its 3 or 7 nearest neighbors (hence $K = 4$ or $K = 8$) to obtain the MMSD estimator of the local subspace. The mean square distance between $\mathcal{R}\left(\hat{\mathbf{H}}_\ell\right)$ and $\mathcal{R}\left(\bar{\mathbf{H}}\right)$, MSD $\left(\hat{\mathbf{H}}_\ell, \bar{\mathbf{H}}\right)$ is then determined to evaluate how close are the local subspace and the global subspace. The results are shown in Figures 10 to 13: for comparison purposes, we display in this figure the result obtained with the SVD, the SMT and the method of [17] which assumes a Bingham prior distribution for $\mathbf{H}$. Figures 10-11 about the Moffett ilage show that a local SVD or SMT would predict rather large differences between the local subspaces and $\bar{\mathbf{H}}$, especially for pixels in the lake area. However, it cannot be concluded that $\bar{\mathbf{H}}$ does not apply for most of the image since, with $K = 4$, the subspace estimated by the SVD may not be very accurate. In contrast, the Bayesian CS-based MMSD estimator shows that $\bar{\mathbf{H}}$ is a rather accurate subspace for the whole image (especially on the lake), except for the pixels along the transition between lake and coastal area. This seems logical as non-linear mixing effects are more likely to occur along the shore, while the linear model is likely to apply well elsewhere. Observe also that $K = 8$ results in a better contrast than $K = 4$. As for Cuprite, Figures 12-13 allow several areas to be clearly identified. From the abundance maps reported in [30], $\bar{\mathbf{H}}$ seems to be far from the estimated local subspaces in parts of the image where kaolinite and muscovite interact. In summary, the MMSD estimator is able to reveal the zones of the image where departure from the linear model might occur. Finally, we note that it is not intuitive to set of value for κ: the values $\kappa = 100$ and $\kappa = 300$ do not have a real meaning and lead to different interpretations of the image. It is much easier to set a value for $\theta_{\max}$, a significant advantage of the CS-based model compared to the method of [17]. However the latter is computationally less intensive. As a final comment, we would like to point out that the computational complexity of the present MMSD-CS method could be prohibitive in large dimensional problems (N large), for which more computationally efficient algorithms, such as the sparse matrix transform of [8], should be favored.

5 Conclusions

In this paper, we considered the problem of subspace estimation from a possibly very limited number of snapshots under the assumption that some prior knowledge about the subspace is available. A Bayesian statistical model was formulated to account for this situation, based on the CS decomposition of the semi-orthogonal matrix $\mathbf{H}$ whose columns span the subspace of interest. This model was shown to rely on rather mild assumptions and, moreover, these assumptions involve meaningful and intuitively appealing quantities, namely the angles between the prior subspace $\bar{\mathbf{H}}$ and the true subspace $\mathbf{H}$. The minimum mean-square distance estimator was implemented through a Gibbs sampling scheme. It was shown to provide accurate estimates, in particular in the low SNR or low sample support regimes. The estimator was also successfully applied to real hyperspectral data, demonstrating its ability to reveal the limits of linear mixing models.

A Sampling from the matrix BMF distribution on $\mathcal{O}(p)$

In this appendix we provide details on how to sample from the Bingham von Mises Fisher distribution on the orthogonal group $\mathcal{O}(p)$ and the derivations below follow those of [23]. The main difference between the Stiefel manifold $\mathcal{S}_{p,N}$ and the orthogonal group lies in the number of columns that are sampled at a time. In $\mathcal{S}_{p,N}$, a random BMF distributed matrix is obtained by successively sampling each column. Since each column is orthogonal to all $p-1$ other columns, it belongs to an $N-p+1$ dimensional subspace and, hence, one needs to draw the length $(N-p+1)$ vector of coordinates in this subspace from a vector BMF distribution. In the case where one needs to sample on the orthogonal group ($N = p$), such a per-column sampling scheme is meaningless as each column should be orthogonal to $p-1$ vectors in a p-dimensional space, and hence there is no degree of freedom left,

only the orientation (sign) of the column vector could be sampled. Therefore, the case of $\mathcal{O}(p)$ requires a special treatment. From the discussion above, one cannot draw one column at a time. Rather, we need to sample successively (at least) two columns which, as will be clarified below, will amount to sample from a matrix BMF distribution on $\mathcal{O}(2)$. Let $\mathbf{X} \in \mathcal{O}(p)$ be a random BMF distributed matrix $\mathbf{X} \sim \text{BMF}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ whose density is given by

$$p(\mathbf{X}) \propto \text{etr}\left\{\mathbf{C}^T\mathbf{X} + \mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{X}\right\} \tag{18}$$

where $\mathbf{A}$ is a symmetric $p \times p$ matrix, $\mathbf{B}$ is a $p \times p$ diagonal matrix and $\mathbf{C}$ is an arbitrary $p \times p$ matrix. In the sequel we let $\mathbf{X}_{[i,j]}$ denote the $p \times 2$ matrix formed from the i-th and j-th columns of $\mathbf{X}$. Accordingly, $\mathbf{X}_{-[i,j]}$ stands for the matrix $\mathbf{X}$ where the i-th and j-th columns have been removed. Finally, we let $\mathbf{X}^{\perp}_{-[i,j]}$ be an orthogonal $p \times 2$ matrix which spans the subspace orthogonal to $\mathcal{R}\left(\mathbf{X}_{-[i,j]}\right)$. From these previous definitions, it follows that $\mathbf{X}_{[i,j]} = \mathbf{X}^{\perp}_{-[i,j]}\mathbf{Y}$ where $\mathbf{Y} \in \mathcal{O}(2)$. As shown in [32], the conditional density of $\mathbf{Y}$ given $\mathbf{X}_{-[i,j]}$ is

$$p\left(\mathbf{Y}|\mathbf{X}_{-[i,j]}\right) \propto \text{etr}\left\{\mathbf{C}^T_{[i,j]}\mathbf{X}^{\perp}_{-[i,j]}\mathbf{Y} + \mathbf{B}_{[i,j]}\mathbf{Y}^T\left(\mathbf{X}^{\perp}_{-[i,j]}\right)^T\mathbf{A}\mathbf{X}^{\perp}_{-[i,j]}\mathbf{Y}\right\} \tag{19}$$

where $\mathbf{B}_{[i,j]} = \text{diag}\left(B(i,i), B(j,j)\right)$. Let $\left(\mathbf{X}^{\perp}_{-[i,j]}\right)^T\mathbf{A}\mathbf{X}^{\perp}_{-[i,j]} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$ be the spectral decomposition of $\left(\mathbf{X}^{\perp}_{-[i,j]}\right)^T\mathbf{A}\mathbf{X}^{\perp}_{-[i,j]}$ and let $\mathbf{Z} = \mathbf{E}^T\mathbf{Y}$, $\mathbf{D} = \mathbf{E}^T\left(\mathbf{X}^{\perp}_{-[i,j]}\right)^T\mathbf{C}_{[i,j]}$. Then, we have

$$p\left(\mathbf{Z}|\mathbf{X}_{-[i,j]}\right) \propto \text{etr}\left\{\mathbf{D}^T\mathbf{Z} + \mathbf{B}_{[i,j]}\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z}\right\}. \tag{20}$$

Now, since $\mathbf{Z} \in \mathcal{O}(2)$, it is necessarily of the form

$$\mathbf{Z} = \begin{pmatrix} \cos\phi & -\varsigma\sin\phi \\ \sin\phi & \varsigma\cos\phi \end{pmatrix} \tag{21}$$

with $\varsigma = \pm 1$. It follows that (20) can be rewritten as

$$\begin{aligned} p\left(\phi, \varsigma|\mathbf{X}_{-[i,j]}\right) &\propto \exp\left\{(d_{11} + \varsigma d_{22})\cos\phi + (d_{21} - \varsigma d_{12})\sin\phi\right\} \\ &\times \exp\left\{(b_1\lambda_1 + b_2\lambda_2)\cos^2\phi + (b_1\lambda_2 + b_2\lambda_1)\sin^2\phi\right\} \end{aligned} \tag{22}$$

where, for notational convenience, we have set $\mathbf{B}_{[i,j]} = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$, $\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ and $\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$. In order to sample from (22), we first sample from

$$\begin{aligned} p\left(\phi|\varsigma, \mathbf{X}_{-[i,j]}\right) &\propto p\left(\phi, \varsigma = 1|\mathbf{X}_{-[i,j]}\right) + p\left(\phi, \varsigma = -1|\mathbf{X}_{-[i,j]}\right) \\ &\propto \exp\left\{d_{11}\cos\phi + d_{21}\sin\phi\right\}\cosh\left(d_{22}\cos\phi - d_{12}\sin\phi\right) \\ &\times \exp\left\{(b_1\lambda_1 + b_2\lambda_2)\cos^2\phi + (b_1\lambda_2 + b_2\lambda_1)\sin^2\phi\right\}. \end{aligned} \tag{23}$$

As the distribution in (23) is rather complicated, the simplest way is to approximate $p\left(\phi|\varsigma, \mathbf{X}_{-[i,j]}\right)$ on a grid of $[0, 2\pi]$ and to sample according to the obtained probabilities. Once ϕ is obtained, we next sample $\varsigma \in \{-1, +1\}$ with probabilities proportional to $\exp\left\{-d_{22}\cos\phi + d_{12}\sin\phi\right\}$ and $\exp\left\{d_{22}\cos\phi - d_{12}\sin\phi\right\}$. The previous scheme allows to sample a pair of columns (i, j). In order to sample the entire matrix, one can draw columns i and j -where j is chosen randomly in $\{1, \cdots, i-1, i+1, \cdots, p\}$- for $i = 1, \cdots, p$.

References

[1] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Reading, MA: Addison Wesley, 1991.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[3] R. Kumaresan and D. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions Acoustics Speech Signal Processing*, vol. 30, no. 6, pp. 833–840, December 1982.

[4] ——, "Estimating the angles of arrival of multiple plane waves," *IEEE Transactions Aerospace Electronic Systems*, vol. 19, no. 1, pp. 134–139, January 1983.

[5] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood and Cramér-Rao bound," *IEEE Transactions Acoustics Speech Signal Processing*, vol. 37, no. 5, pp. 720–741, May 1989.

[6] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, February 2004.

[7] T. L. Marzetta, G. H. Tucci, and S. H. Simon, "A random matrix theoretic approach to handling singular covariance estimates," *IEEE Transactions Information Theory*, vol. 57, no. 9, pp. 6256–6271, September 2011.

[8] G. Cao, L. R. Bachega, and C. A. Bouman, "The sparse matrix transform for covariance estimation and analysis of high dimensional signals," *IEEE Transactions Image Processing*, vol. 20, no. 3, pp. 625–640, March 2011.

[9] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates," *IEEE Transactions Information Theory*, vol. 54, no. 11, pp. 5113–5129, November 2008.

[10] J. Thomas, L. Scharf, and D. Tufts, "The probability of a subspace swap in the SVD," *IEEE Transactions Signal Processing*, vol. 43, no. 3, pp. 730–736, March 1995.

[11] M. Hawkes, A. Nehorai, and P. Stoica, "Performance breakdown of subspace-based methods: prediction and cure," in *Proceedings ICASSP*, May 2001, pp. 4005–4008.

[12] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed., ser. Springer Series in Statistics. New York: Springer Verlag, 2010.

[13] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, October 2007.

[14] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *ArXiv:1103.2221*, March 2011.

[15] ——, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 211, no. 1, pp. 494–521, May 2011.

[16] A. Srivastava, "A Bayesian approach to geometric subspace estimation," *IEEE Transactions Signal Processing*, vol. 48, no. 5, pp. 1390–1400, May 2000.

[17] O. Besson, N. Dobigeon, and J.-Y. Tourneret, "Minimum mean square distance estimation of a subspace," *IEEE Transactions Signal Processing*, vol. 59, no. 12, pp. 5709–5720, December 2011.

[18] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal Matrix Analysis Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[19] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed. Baltimore: John Hopkins University Press, 1996.

[20] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, 1999.

[21] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer Verlag, 2004.

[22] C. P. Robert, *The Bayesian Choice - From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer Verlag, 2007.

[23] P. D. Hoff, "Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, vol. 18, no. 2, pp. 438–456, June 2009.

[24] C.-I Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Kluwer Academic, 2003.

[25] D. Manolakis, C. Siracusa, and G. Shaw, "Hyperspectral subpixel target detection using the linear mixing model," *IEEE Transactions Geoscience Remote Sensing*, vol. 39, no. 7, pp. 1392–409, July 2001.

[26] M. Lewis, V. Jooste, and A. A. de Gasparis, "Discrimination of arid vegetation with airborne multispectral scanner hyperspectral imagery," *IEEE Transactions Geoscience Remote Sensing*, vol. 39, no. 7, pp. 1471–1479, July 2001.

[27] B. Datt, T. R. McVicar, T. G. V. Niel, D. L. B. Jupp, and J. S. Pearlman, "Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes," *IEEE Transactions Geoscience Remote Sensing*, vol. 41, no. 6, pp. 1246–1259, June 2003.

[28] J. Plaza, R. Pérez, A. Plaza, P. Martínez, and D. Valencia, "Mapping oil spills on sea water using spectral mixture analysis of hyperspectral image data," in *Chemical and Biological Standoff Detection III*, J. O. Jensen and J.-M. Thériault, Eds., vol. 5995. SPIE, 2005, pp. 79–86.

[29] N. Keshava and J. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, January 2002.

[30] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint Bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions Signal Processing*, vol. 57, no. 11, pp. 4355–4368, November 2009.

[31] R. N. Clark, G. A. Swayze, and A. Gallagher, "Mapping minerals with imaging spectroscopy, u.s. geological survey," *Office of Mineral Resources Bulletin*, vol. 2039, pp. 141–150, 1993.

[32] Y. Chikuse, *Statistics on special manifolds*. New York: Springer Verlag, 2003.

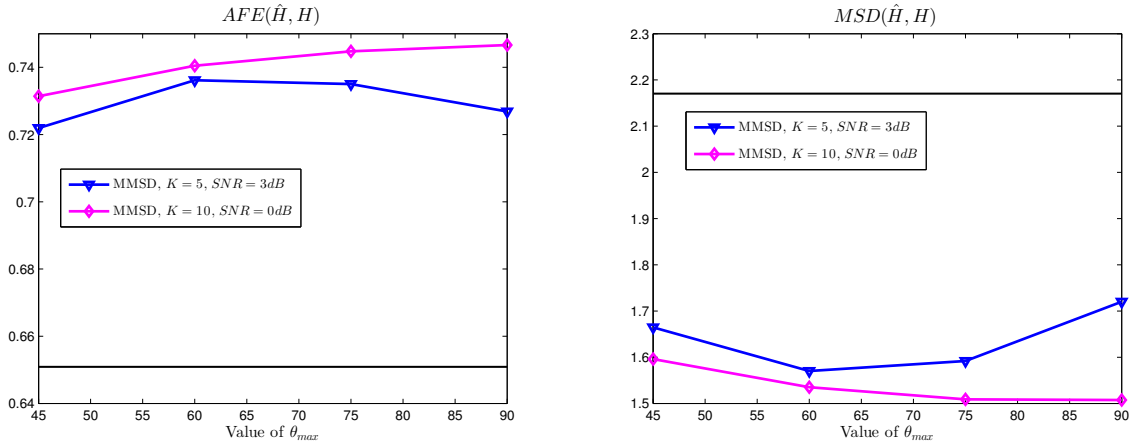

Figure 1: Average fraction of energy and mean-square distance between true and estimated subspaces versus $\theta_{\max}$. $N = 20$, $p = 5$.

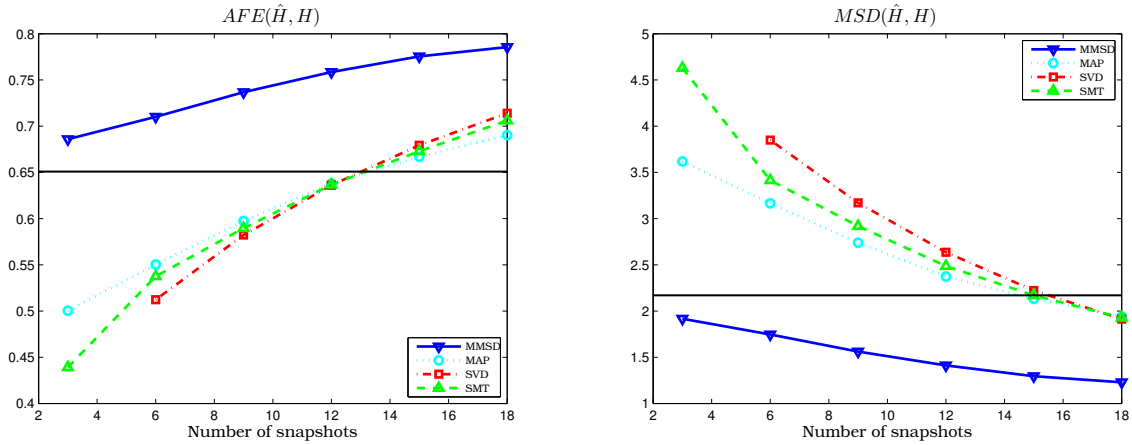

Figure 2: Average fraction of energy and mean-square distance between true and estimated subspaces versus K. $N = 20$, $p = 5$, SNR $= 0$dB and $\theta_{\max} = 60°$.

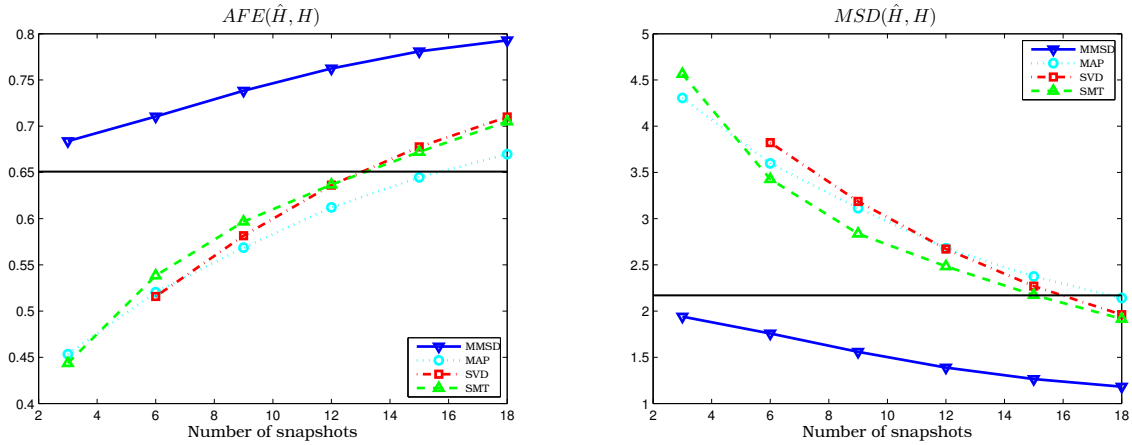

Figure 3: Average fraction of energy and mean-square distance between true and estimated subspaces versus K. $N = 20$, $p = 5$, SNR $= 0$dB and $\theta_{\max} = 75°$.

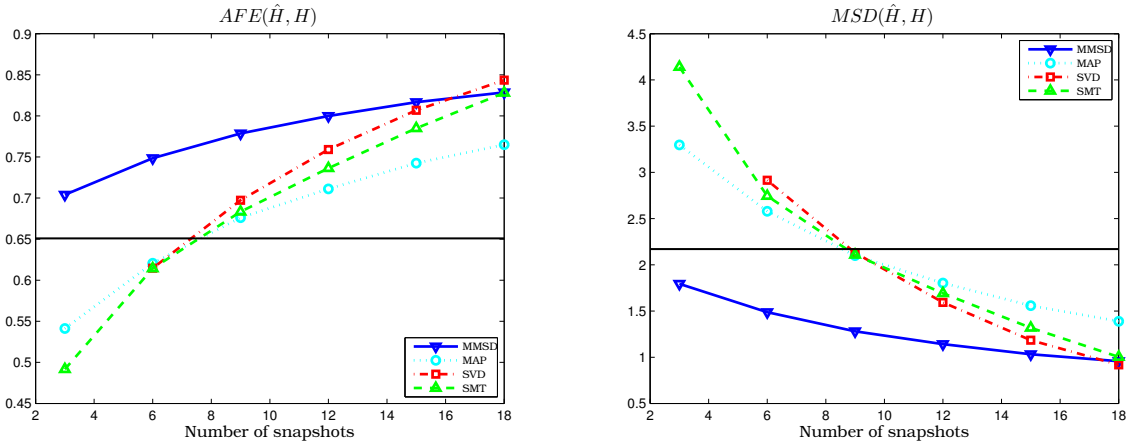

Figure 4: Average fraction of energy and mean-square distance between true and estimated subspaces versus K. $N = 20$, $p = 5$, SNR $= 3$dB and $\theta_{\max} = 60°$.

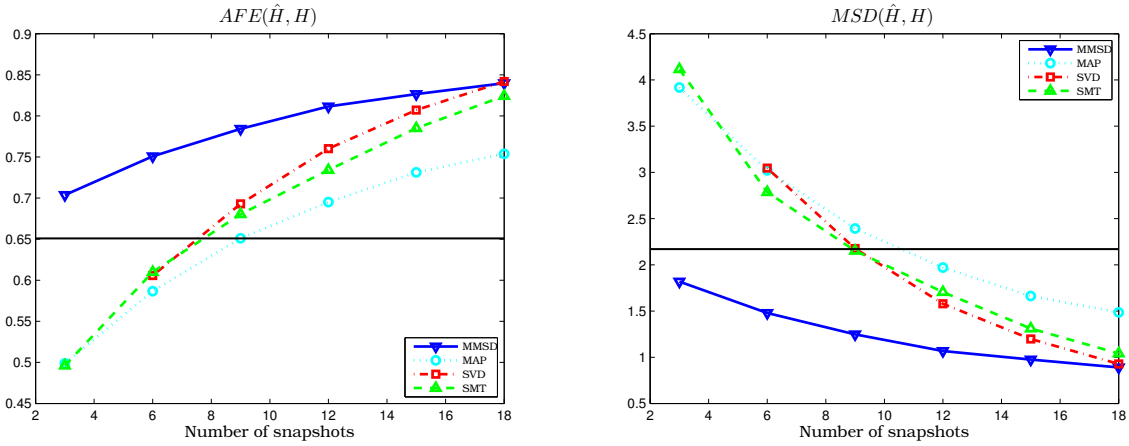

Figure 5: Average fraction of energy and mean-square distance between true and estimated subspaces versus K. $N = 20$, $p = 5$, SNR $= 3$dB and $\theta_{\max} = 75°$.

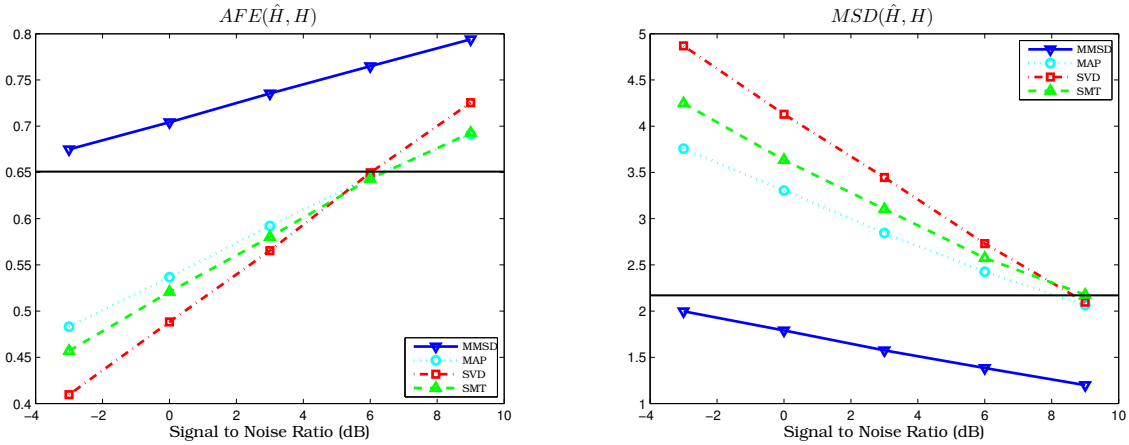

Figure 6: Average fraction of energy and mean-square distance between true and estimated subspaces versus SNR. $N = 20$, $p = 5$, $K = 5$ and $\theta_{\max} = 60°$.

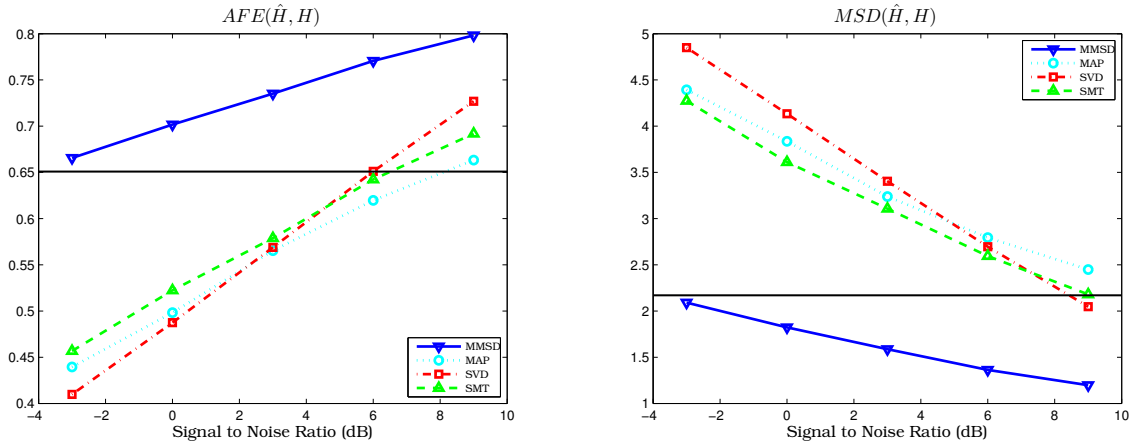

Figure 7: Average fraction of energy and mean-square distance between true and estimated subspaces versus SNR. $N = 20$, $p = 5$, $K = 5$ and $\theta_{\max} = 75°$.

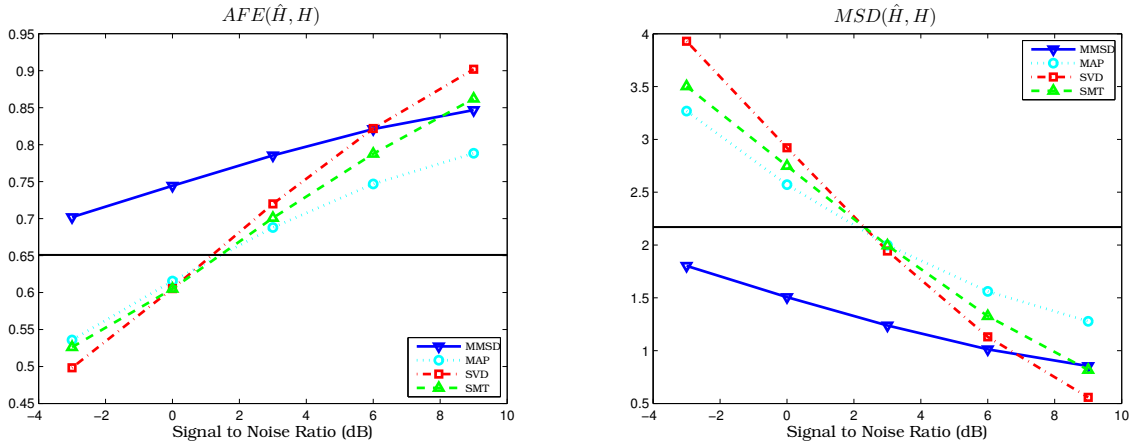

Figure 8: Average fraction of energy and mean-square distance between true and estimated subspaces versus SNR. $N = 20$, $p = 5$, $K = 10$ and $\theta_{\max} = 60°$.

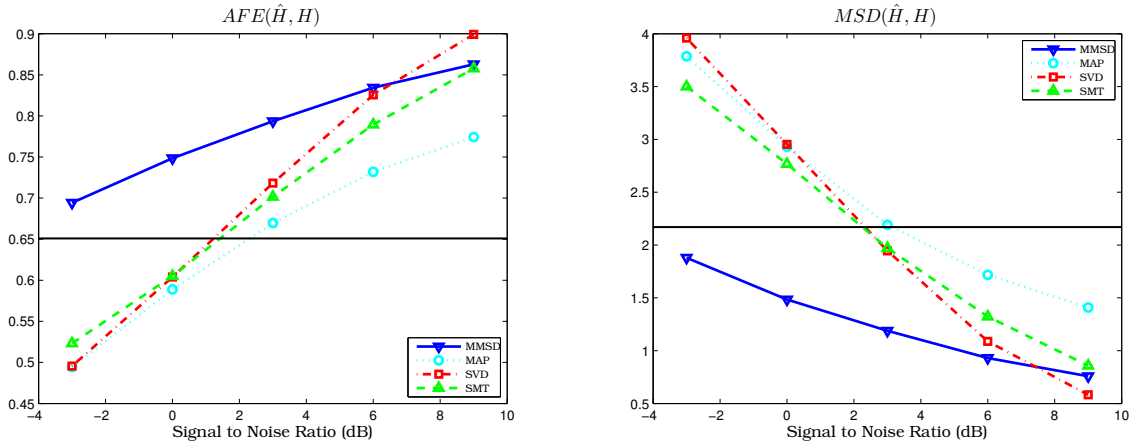

Figure 9: Average fraction of energy and mean-square distance between true and estimated subspaces versus SNR. $N = 20$, $p = 5$, $K = 10$ and $\theta_{\max} = 75°$.

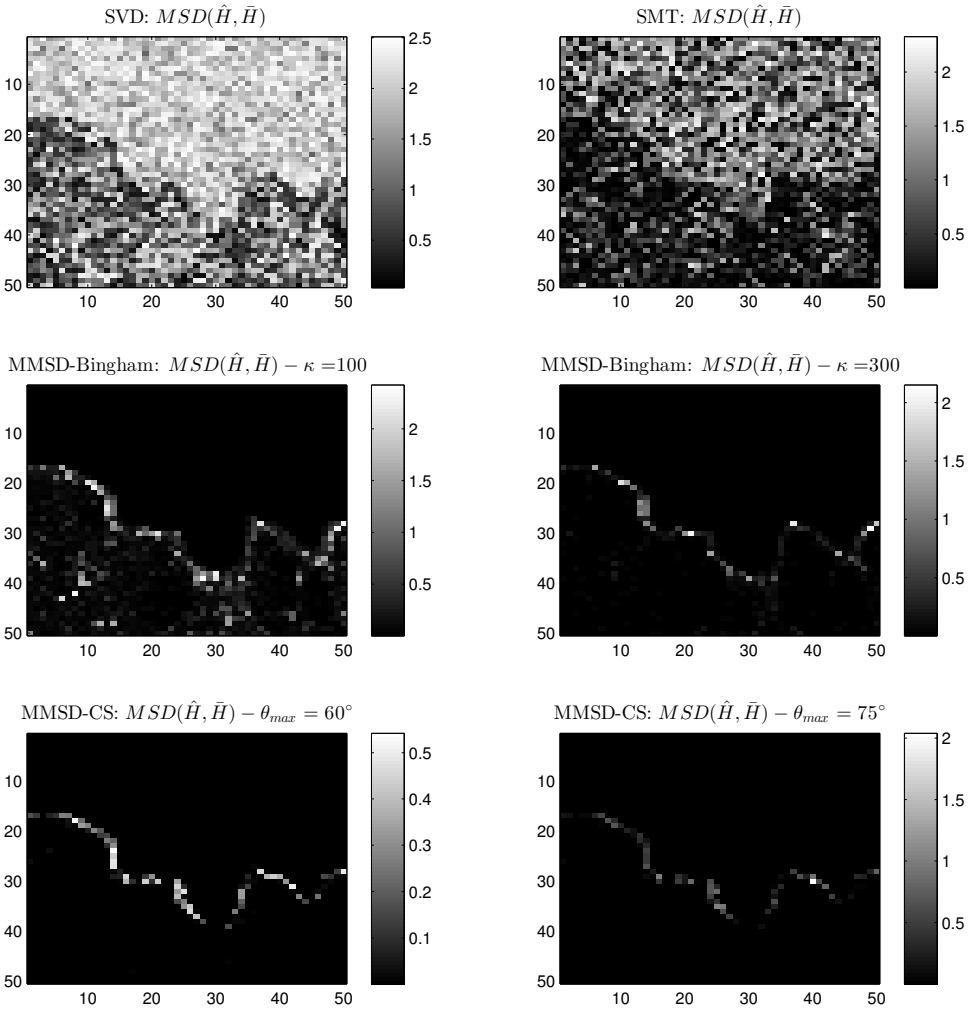

Figure 10: Moffett image. MSD $\left(\hat{\mathbf{H}}_{\ell},\bar{\mathbf{H}}\right)$. $N=183$, $p=2$, $K=4$.

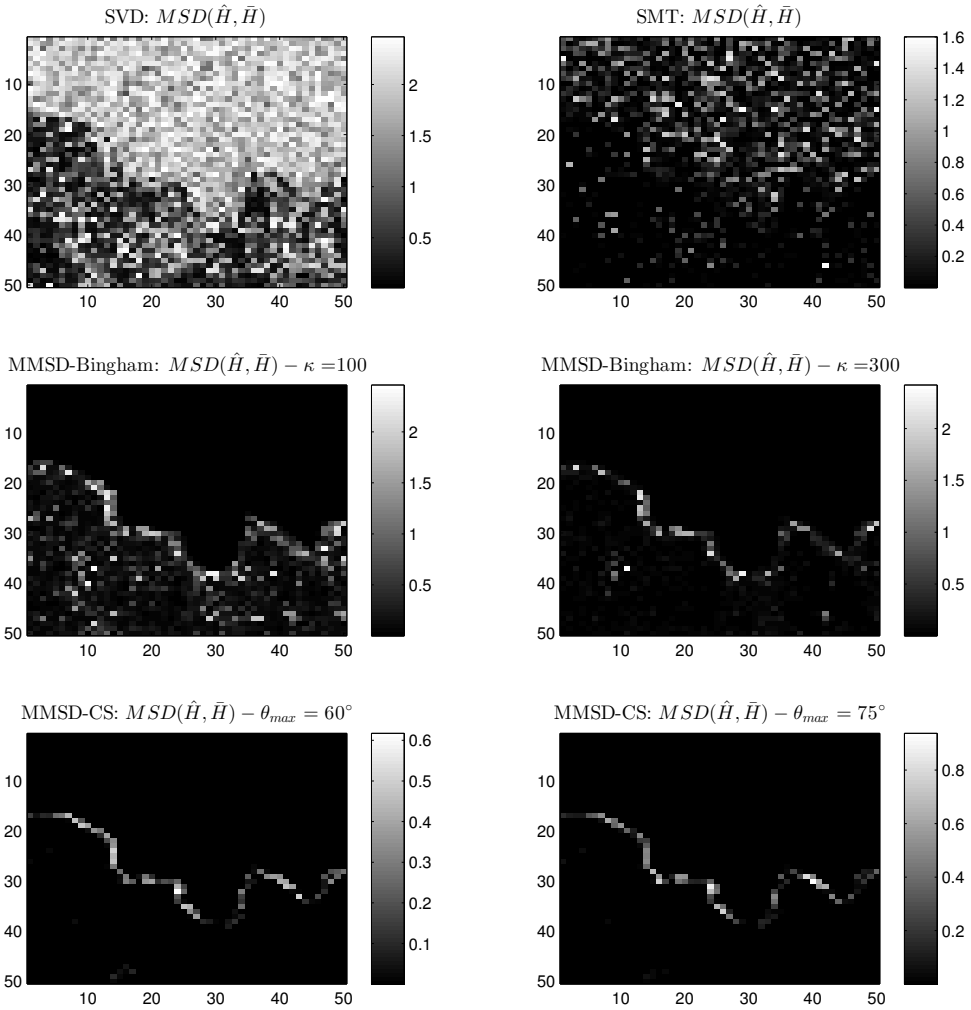

Figure 11: Moffett image. MSD $\left(\hat{\mathbf{H}}_{\ell},\bar{\mathbf{H}}\right)$. $N=183$, $p=2$, $K=8$.

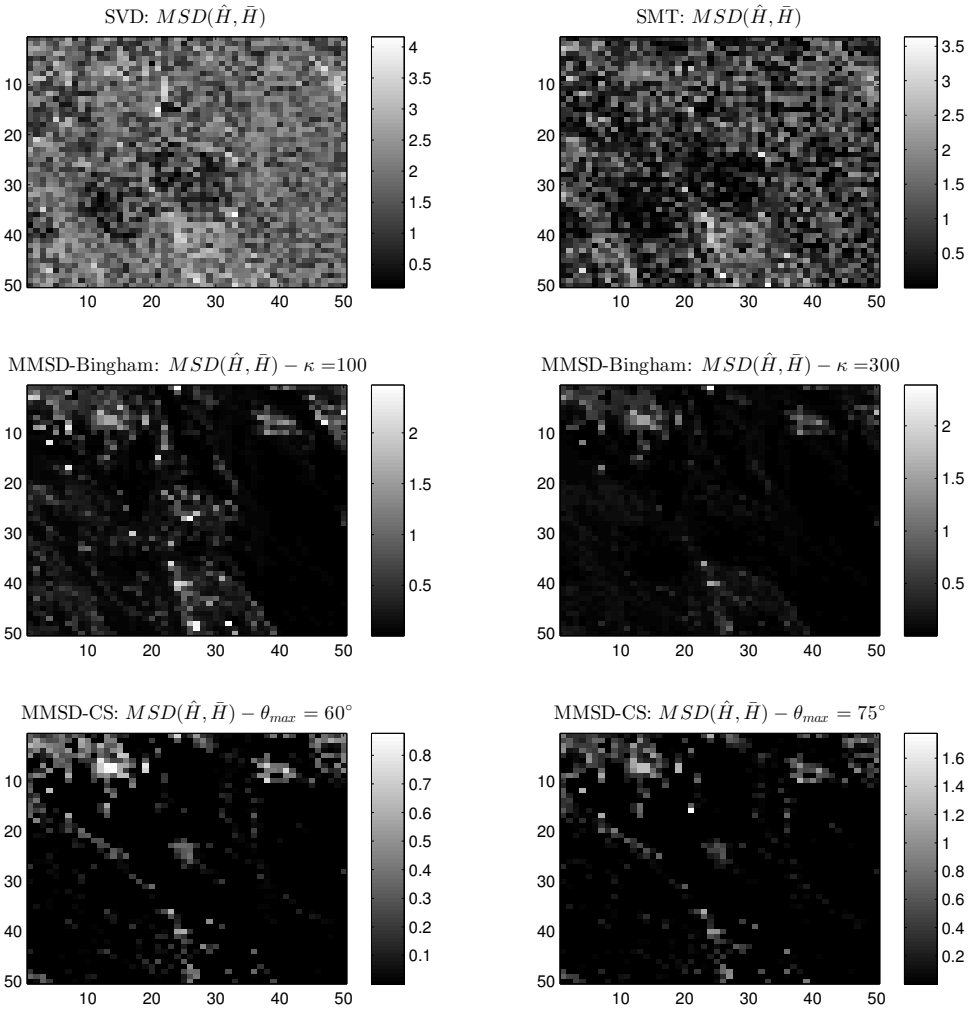

Figure 12: Cuprite image. MSD $\left(\hat{\mathbf{H}}_{\ell},\bar{\mathbf{H}}\right)$. $N=189$, $p=2$, $K=4$.

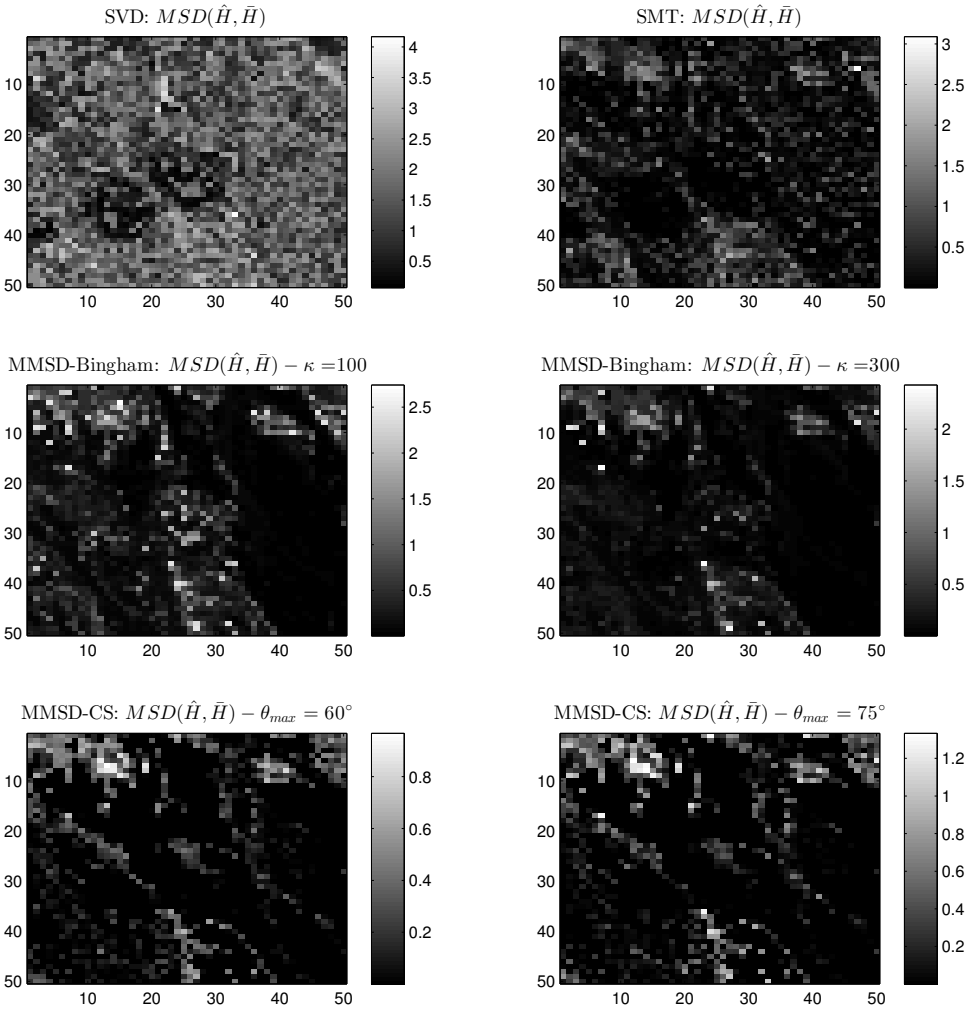

Figure 13: Cuprite image. MSD $\left(\hat{\mathbf{H}}_{\ell},\bar{\mathbf{H}}\right)$. $N=189$, $p=2$, $K=8$.