Minimum mean square distance estimation of a subspace

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

Technical report - 14 June 2011

Abstract

We consider the problem of subspace estimation in a Bayesian setting. Since we are operating in the Grassmann manifold, the usual approach which consists of minimizing the mean square error (MSE) between the true subspace $\boldsymbol{U}$ and its estimate $\hat{\boldsymbol{U}}$ may not be adequate as the MSE is not the natural metric in the Grassmann manifold $G_{N,p}$, i.e., the set of p-dimensional subspaces in $\mathbb{R}^N$. As an alternative, we propose to carry out subspace estimation by minimizing the mean square distance between $\boldsymbol{U}$ and its estimate, where the considered distance is a natural metric in the Grassmann manifold, viz. the distance between the projection matrices. We show that the resulting estimator is no longer the posterior mean of $\boldsymbol{U}$ but entails computing the principal eigenvectors of the posterior mean of $\boldsymbol{U}\boldsymbol{U}^T$. Derivation of the minimum mean square distance (MMSD) estimator is carried out in a few illustrative examples including a linear Gaussian model for the data and Bingham or von Mises Fisher prior distributions for $\boldsymbol{U}$. In all scenarios, posterior distributions are derived and the MMSD estimator is obtained either analytically or implemented via a Markov chain Monte Carlo simulation method. The method is shown to provide accurate estimates even when the number of samples is lower than the dimension of $\boldsymbol{U}$. An application to hyperspectral imagery is finally investigated.

*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

In many signal processing applications, the signals of interest do not span the entire observation space and a relevant and frequently used assumption is that they evolve in a low-dimensional subspace [1]. Subspace modeling is accurate when the signals consist of a linear combination of p modes in an N-dimensional space, and constitute a good approximation for example when the signal covariance matrix is close to rank-deficient. As a consequence, subspace estimation plays a central role in recovering these signals with maximum accuracy. Frequency estimation of complex exponential signals and directions of arrival (DoA) of signals using an array of sensors are examples of well-known area where obtaining accurate subspace estimates constitutes a crucial task. In the latter case, the data matrix can be written as $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{S}+\boldsymbol{N}$ where the columns of the $N \times p$ matrix $\boldsymbol{A}(\boldsymbol{\theta})$ contain the p steering vectors of the sources of interest, and $\boldsymbol{\theta}$ corresponds to the vector of their DoA. In a noise-free environment, $\boldsymbol{Y}$ is rank-deficient and its column space is spanned by the steering vectors, from which DoA estimation can be performed. Therefore, a key issue is to first identify the dominant subspace of $\boldsymbol{Y}$ in order then to infer the DoA. This is the essence of most subspace-based methods, such as MUSIC [2], ESPRIT [3] or Min-Norm [4,5]. These methods have gained much popularity because they represent computationally interesting alternatives to the maximum likelihood (ML) estimator [6]. Moreover, they yield equivalent performance in the asymptotic regime, viz large number of snapshots or high signal to noise ratio (SNR) [7, 8]. An ubiquitous tool to compute the dominant subspace is the singular value decomposition (SVD) of the data matrix or, equivalently, the eigenvalue decomposition of the sample covariance matrix. Observe also that the SVD emerges naturally as the maximum likelihood estimator of the range space $\mathcal{R}(\boldsymbol{U})$ of $\boldsymbol{U}$ in the classical model $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{S} + \boldsymbol{N}$, where $\boldsymbol{Y}$ stands for the $N \times K$ observation matrix, $\boldsymbol{U}$ is the (deterministic) $N \times p$ matrix, with $p < N$, whose columns span the p-dimensional subspace of interest, $\boldsymbol{S}$ is the $p \times K$ (deterministic) waveform matrix and $\boldsymbol{N}$ is the additive noise. The latter model is relevant e.g., in hyperspectral imagery, see Section 5 for further discussion.

When the SNR is high or the number of snapshots is large, the SVD yields quasi-optimal estimates. However, the SVD can incur some performance loss in two main cases, namely when the SNR is very low and thereof the probability of a subspace swap or subspace leakage is high [9–12]. A second case occurs when the number of samples K is small, typically of the order of N or, even less, of the order of p. In order to overcome these limitations, especially the case of low sample support, some interesting alternatives have been proposed, such as [13], where accurate estimates of the eigenvalues and eigenvectors are derived, based on random matrix theory. However, the sample support considered there is still higher than what we consider herein. More precisely, we are mostly interested in the case where the number of snapshots is very small and may be less than the subspace dimension p: in this case, $\boldsymbol{Y}$ is at most of rank K and information is lacking about how to complement $\mathcal{R}(\boldsymbol{Y})$ in order to estimate $\mathcal{R}(\boldsymbol{U})$. This case along with the low SNR case are those of most interest to us and, hence, we need to turn to a different methodology.

In such situations, a Bayesian approach might be helpful as it enables one to assist estimation by providing some statistical information about $\boldsymbol{U}$. We investigate such an approach herein and assign to the unknown matrix $\boldsymbol{U}$ an appropriate prior distribution, taking into account the specific structure of $\boldsymbol{U}$. The paper is organized as follows. In section 2, we propose an approach based on minimizing a natural distance on the Grassmann manifold, which yields a new estimator of $\boldsymbol{U}$. The theory is illustrated in section 3 where the new estimator is derived for some specific examples. In section 4 its performance is assessed through numerical simulations, and compared with conventional approaches. Section 5 studies an application to the analysis of interactions between pure materials contained in hyperspectral images.

2 Minimum mean square distance estimation

As explained above, we adopt a Bayesian framework for estimation of the unknown matrix $\boldsymbol{U}$ and consider an alternative to the conventional minimum mean square error (MMSE) estimator, which is usually regarded as the chief systematic approach under the Bayesian umbrella [14]. Let us consider that we wish to estimate the range space of $\boldsymbol{U}$ from the joint distribution $p(\boldsymbol{Y}, \boldsymbol{U})$ where $\boldsymbol{Y}$ stands for the available data matrix. Usually, one is not interested in $\boldsymbol{U}$ *per se* but rather in its range space $\mathcal{R}(\boldsymbol{U})$, and thus we are operating in the Grassmann manifold $G_{N,p}$, i.e., the set of p-dimensional subspaces in $\mathbb{R}^N$ [15]. It is thus natural to wonder whether the MMSE estimator is suitable in $G_{N,p}$. The MMSE estimator $\hat{\boldsymbol{\theta}}$ of a vector $\boldsymbol{\theta}$ minimizes the average squared Euclidean distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$, i.e., $\mathsf{E}\left\{\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2\right\}$. Despite the fact that this distance is natural in an Euclidean space, it may not be the more natural metric in $G_{N,p}$. In fact, the natural distance between two subspaces $\mathcal{R}(\boldsymbol{U}_1)$ and $\mathcal{R}(\boldsymbol{U}_2)$ is given by $\left(\sum_{k=1}^{p} \theta_k^2\right)^{1/2}$ [15] where θ_k are the principal angles between these subspaces, which can be obtained by SVD of $\boldsymbol{U}_2^T \boldsymbol{U}_1$ where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ denote orthonormal bases for these subspaces [16]. The SVD of $\boldsymbol{U}_2^T \boldsymbol{U}_1$ is defined as $\boldsymbol{U}_2^T \boldsymbol{U}_1 = \boldsymbol{X} \operatorname{diag}(\cos\theta_1, \cdots, \cos\theta_p) \boldsymbol{Z}^T$, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are two $p \times p$ unitary matrices. Therefore, it seems more adequate, rather than minimizing $\left\|\hat{\boldsymbol{U}} - \boldsymbol{U}\right\|_F^2$ as the MMSE estimator does, to minimize the natural distance between the subspaces spanned by $\hat{\boldsymbol{U}}$ and $\boldsymbol{U}$. Although this is the most intuitively appealing method, it faces the drawback that the cosines of the angles and not the angles themselves emerge naturally from the SVD. Therefore, for the sake of practicality, we consider minimizing the sum of the squared sine of the angles between $\hat{\boldsymbol{U}}$ and $\boldsymbol{U}$. As argued in [15, 16], this cost function is natural in the Grassmann manifold since it corresponds to the Frobenius norm of the difference between the projection matrices on the two subspaces, viz $\sum_{k=1}^{p} \sin^2\theta_k = \left\|\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T - \boldsymbol{U}\boldsymbol{U}^T\right\|_F^2 \triangleq d^2\left(\hat{\boldsymbol{U}}, \boldsymbol{U}\right)$. It should be mentioned that our approach follows along the same principles as in [17] where a Bayesian framework is proposed for subspace estimation, and where the author considers minimizing $d\left(\hat{\boldsymbol{U}}, \boldsymbol{U}\right)$. Hence the theory presented in this section is similar to that of [17], only the parameterization of the problem in [17] being slightly different from ours. The main differences compared to [17] lie in the prior distributions and signal models used, as well as in the implementation of the MMSD estimator, see the next section for details.

Given that $d^2\left(\hat{\boldsymbol{U}}, \boldsymbol{U}\right) = 2\left(p - \operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\}\right)$, we define the minimum mean-square distance (MMSD) estimator of $\boldsymbol{U}$ as

$$\hat{\boldsymbol{U}}_{\text{mmsd}} = \arg\max_{\hat{\boldsymbol{U}}} \mathsf{E}\left\{\operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\}\right\}. \tag{1}$$

Since

$$\mathsf{E}\left\{\operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\}\right\} = \int \left[\int \operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\} p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}\right] p(\boldsymbol{Y})\, d\boldsymbol{Y} \tag{2}$$

it follows that

$$\begin{aligned}\hat{\boldsymbol{U}}_{\text{mmsd}} &= \arg\max_{\hat{\boldsymbol{U}}} \int \operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\} p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U} \\ &= \arg\max_{\hat{\boldsymbol{U}}} \operatorname{Tr}\left\{\hat{\boldsymbol{U}}^T \left[\int \boldsymbol{U}\boldsymbol{U}^T p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}\right] \hat{\boldsymbol{U}}\right\}. \end{aligned} \tag{3}$$

Therefore, the MMSD estimate of the subspace spanned by $\boldsymbol{U}$ is given by the p largest eigenvectors of the matrix $\int \boldsymbol{U}\boldsymbol{U}^T p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}$, which we denote as

$$\hat{\boldsymbol{U}}_{\text{mmsd}} = \mathcal{P}_p\left\{\int \boldsymbol{U}\boldsymbol{U}^T p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}\right\}. \tag{4}$$

In other words, MMSD estimation amounts to find the best rank-p approximation to the posterior mean of the projection matrix $\boldsymbol{U}\boldsymbol{U}^T$ on $\mathcal{R}(\boldsymbol{U})$. For notational convenience, let us denote $\boldsymbol{M}(\boldsymbol{Y}) = \int \boldsymbol{U}\boldsymbol{U}^T p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}$. Except for a few cases where this matrix can be derived in closed-form (an example is given in the next section), there usually does not exist any analytical expression for $\boldsymbol{M}(\boldsymbol{Y})$. In such situation, an efficient way to approximate the matrix $\boldsymbol{M}(\boldsymbol{Y})$ is to use a Markov chain Monte Carlo (MCMC) simulation method whose goal is to generate random matrices $\boldsymbol{U}$ drawn from the posterior distribution $p(\boldsymbol{U}|\boldsymbol{Y})$, and to approximate the integral in (4) by a finite sum. This aspect will be further elaborated in the next section. Let $\boldsymbol{M}(\boldsymbol{Y}) = \boldsymbol{U}_M(\boldsymbol{Y})\boldsymbol{L}_M(\boldsymbol{Y})\boldsymbol{U}_M^T(\boldsymbol{Y})$ denote the eigenvalue decomposition of $\boldsymbol{M}(\boldsymbol{Y})$ with $\boldsymbol{L}_M(\boldsymbol{Y}) = \mathrm{diag}\,(\ell_1(\boldsymbol{Y}), \ell_2(\boldsymbol{Y}), \cdots, \ell_N(\boldsymbol{Y}))$ and $\ell_1(\boldsymbol{Y}) \geq \ell_2(\boldsymbol{Y}) \geq \cdots \geq \ell_N(\boldsymbol{Y})$. Then the average distance between $\hat{\boldsymbol{U}}_{\mathrm{mmsd}}$ and $\boldsymbol{U}$ is given by

$$\begin{aligned}
\mathsf{E}\left\{d^2\left(\hat{\boldsymbol{U}}_{\mathrm{mmsd}}, \boldsymbol{U}\right)\right\} &= 2p - 2\int\left[\int \mathrm{Tr}\left\{\hat{\boldsymbol{U}}_{\mathrm{mmsd}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}_{\mathrm{mmsd}}\right\} p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}\right] p(\boldsymbol{Y})\, d\boldsymbol{Y} \\
&= 2p - 2\int \mathrm{Tr}\left\{\hat{\boldsymbol{U}}_{\mathrm{mmsd}}^T \boldsymbol{M}(\boldsymbol{Y}) \hat{\boldsymbol{U}}_{\mathrm{mmsd}}\right\} p(\boldsymbol{Y})\, d\boldsymbol{Y} \\
&= 2p - 2\sum_{k=1}^{p}\int \ell_k(\boldsymbol{Y}) p(\boldsymbol{Y})\, d\boldsymbol{Y}. \qquad (5)
\end{aligned}$$

The latter expression constitutes a lower bound on $\mathsf{E}\left\{d^2\left(\hat{\boldsymbol{U}}, \boldsymbol{U}\right)\right\}$ and is referred to as the Hilbert-Schmidt bound in [17, 18]. As indicated in these references, and similarly to $\boldsymbol{M}(\boldsymbol{Y})$, this lower bound may be difficult to obtain analytically.

The MMSD approach can be extended to the mixed case where, in addition to $\boldsymbol{U}$, a parameter vector $\boldsymbol{\theta}$ which can take arbitrary values in $\mathbb{R}^q$ needs to be estimated jointly with $\boldsymbol{U}$. Under such circumstances, one can estimate $\boldsymbol{U}$ and $\boldsymbol{\theta}$ as

$$\left(\hat{\boldsymbol{U}}_{\mathrm{mmsd}}, \hat{\boldsymbol{\theta}}_{\mathrm{mmsd}}\right) = \arg\min_{\hat{\boldsymbol{U}}, \hat{\boldsymbol{\theta}}} \mathsf{E}\left\{-\mathrm{Tr}\left\{\hat{\boldsymbol{U}}^T \boldsymbol{U}\boldsymbol{U}^T \hat{\boldsymbol{U}}\right\} + \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right\}. \qquad (6)$$

Doing so, the MMSD estimator of $\boldsymbol{U}$ is still be given by (4) while the MMSD and MMSE estimators of $\boldsymbol{\theta}$ coincide.

Remark 1. The MMSD approach differs from an MMSE approach which would entail calculating the posterior mean of $\boldsymbol{U}$, viz $\int \boldsymbol{U} p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}$. Note that the latter may not be meaningful, in particular when the posterior distribution $p(\boldsymbol{U}|\boldsymbol{Y})$ depends on $\boldsymbol{U}$ only through $\boldsymbol{U}\boldsymbol{U}^T$, see next section for an example. In such a case, post-multiplication of $\boldsymbol{U}$ by any $p \times p$ unitary matrix $\boldsymbol{Q}$ yields the same value of $p(\boldsymbol{U}|\boldsymbol{Y})$. Therefore averaging $\boldsymbol{U}$ over $p(\boldsymbol{U}|\boldsymbol{Y})$ does not make sense while computing (4) is relevant. On the other hand, if $p(\boldsymbol{U}|\boldsymbol{Y})$ depends on $\boldsymbol{U}$ directly, then computing the posterior mean of $\boldsymbol{U}$ can be investigated: an example where this situation occurs will be presented in the next section. As a final comment, observe that $\int \boldsymbol{U} p(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}$ is not necessarily unitary but its range space can be used to estimate $\mathcal{R}(\boldsymbol{U})$.

Remark 2. We open a parenthesis here regarding the framework of this paper. First of all, we do not deal herein with sequential estimation in a Bayesian framework as can be the case e.g., in [17, 19]. In contrast, we consider batch algorithms where a given number of snapshots is used to produce an estimate of $\boldsymbol{U}$. A second important notice is that the MMSD estimator of this paper originally stems from the minimization of a certain distance over the Grassmann manifold or the Stiefel manifold (the set of $N \times p$ matrices $\boldsymbol{U}$ such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$). Optimization over these manifolds has received considerable attention recently, see the excellent tutorial paper by Edelman *et al.* [15] as well as [20, 21], and [22–24] for signal processing applications. More generally, minimization over special manifolds (see e.g., [25] for optimization over the special

linear group) is attracting interest as it enables the underlying geometry of the problem to be taken into account. In the aforementioned references, the focus is on deriving iterative methods, such as steepest descent, conjugate gradient or Newton methods. In the present paper, the problem addressed is slightly different: we do not consider optimization with respect to an unknown matrix $\boldsymbol{U}$, rather we need to compute the average of $\boldsymbol{U}\boldsymbol{U}^T$ over $p(\boldsymbol{U}|\boldsymbol{Y})$ when $\boldsymbol{U}$ is a random matrix on the Stiefel manifold.

3 Illustration examples

In this section we illustrate the previous theory on some examples, including the conventional linear Gaussian model (conditioned on $\boldsymbol{U}$) and a model involving the eigenvalue decomposition of the data covariance matrix. As a first step, we address the issue of selecting prior distributions for $\boldsymbol{U}$ and then move on to the derivation of the MMSD estimator.

3.1 Prior distributions

A crucial step in any Bayesian estimation scheme consists of selecting the prior distribution for the variables to be estimated. We focus here on distributions on the Stiefel or Grassmann manifold, depending whether we consider the matrix $\boldsymbol{U}$ itself or its range space. There exist only a few distributions on the Stiefel or Grassmann manifolds, the most widely accepted being the Bingham or von Mises Fisher (vMF) distributions [26, 27], which are given respectively by

$$p_{\mathrm{B}}(\boldsymbol{U}) = \frac{1}{{}_1F_1\left(\frac{1}{2}p, \frac{1}{2}N; \boldsymbol{A}\right)} \mathrm{etr}\left\{\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}\right\} \tag{7}$$

$$p_{\mathrm{vMF}}(\boldsymbol{U}) = \frac{1}{{}_0F_1\left(\frac{1}{2}N; \frac{1}{4}\boldsymbol{F}^T\boldsymbol{F}\right)} \mathrm{etr}\left\{\boldsymbol{F}^T\boldsymbol{U}\right\} \tag{8}$$

where $\mathrm{etr}\{.\}$ stands for the exponential of the trace of the matrix between braces, $\boldsymbol{A}$ is an $N \times N$ symmetric matrix, $\boldsymbol{F}$ is an $N \times p$ arbitrary matrix, and ${}_0F_1(a; \boldsymbol{X})$, ${}_1F_1(a, b; \boldsymbol{X})$ are hypergeometric functions of matrix arguments, see e.g., [27] for their definitions. The Bingham and the von Mises Fisher distributions have been proposed in various applications, including meteorology, biology, medicine, image analysis (see [26] and references therein), and recently for modeling of multipath communications channels [28,29] or in shape analysis [30]. We will denote these distributions as $\mathrm{B}(\boldsymbol{A})$ and $\mathrm{vMF}(\boldsymbol{F})$, respectively. Observe that the Bingham distribution depends on $\boldsymbol{U}\boldsymbol{U}^T$ only, and can thus be viewed as a distribution on the Grassmann manifold [26, 27] while the vMF distribution depends on $\boldsymbol{U}$ and is a distribution on the Stiefel manifold. In most applications, one is mostly interested in the projection matrix $\boldsymbol{U}\boldsymbol{U}^T$ and therefore the Bingham distribution appears to be a more natural choice than the vMF distribution. In our case, in order to introduce some knowledge about $\boldsymbol{U}$, we assume that it is "close" to a given subspace spanned by the columns of an orthonormal matrix $\bar{\boldsymbol{U}}$, and hence we consider two possible prior distributions for $\boldsymbol{U}$, namely

$$\pi_{\mathrm{B}}(\boldsymbol{U}) \propto \mathrm{etr}\left\{\kappa\boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U}\right\} \tag{9}$$

$$\pi_{\mathrm{vMF}}(\boldsymbol{U}) \propto \mathrm{etr}\left\{\kappa\boldsymbol{U}^T\bar{\boldsymbol{U}}\right\} \tag{10}$$

where $\propto$ means "proportional to". The matrix $\bar{\boldsymbol{U}}$ reflects our knowledge about the subspace where the signals evolve. This matrix can be obtained from the data itself (see section 5.2 for an example about hyperspectral imagery) or from some available models. For instance, in radar applications, if the clutter subspace is to be estimated, there exists a number of relevant models, including the general covariance matrix model of [31], that can be used to obtain $\bar{\boldsymbol{U}}$.

The distribution in (9) is proportional to the sum of the squared cosine angles between $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$ while $\pi_{\text{vMF}}(\boldsymbol{U})$ is proportional to the sum of the cosine angles between $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$. Note that κ is a concentration parameter: the larger κ the more concentrated around $\bar{\boldsymbol{U}}$ are the subspaces $\boldsymbol{U}$. The difference between the two distributions is the following. In the Bingham distribution only $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$ are close (at least for large values of κ) since $\pi_{\text{B}}(\boldsymbol{U})$ is invariant to post-multiplication of $\boldsymbol{U}$ by any $p \times p$ unitary matrix $\boldsymbol{Q}$. Hence $\boldsymbol{U}$ is not necessarily close to $\bar{\boldsymbol{U}}$. In contrast, under the vMF prior distribution, $\boldsymbol{U}$ and $\bar{\boldsymbol{U}}$ are close. For illustration purposes, Figure 1 displays the average fraction of energy of $\boldsymbol{U}$ in $\mathcal{R}(\bar{\boldsymbol{U}})$ defined as

$$\text{AFE}(\boldsymbol{U}, \bar{\boldsymbol{U}}) = \mathsf{E}\left\{\text{Tr}\left\{\boldsymbol{U}^T \bar{\boldsymbol{U}} \bar{\boldsymbol{U}}^T \boldsymbol{U}\right\}/p\right\}. \tag{11}$$

As can be observed from these figures, both distributions allow the distance between $\boldsymbol{U}$ and $\bar{\boldsymbol{U}}$ to be set in a rather flexible way. Their AFE is shown to be identical for small values of the concentration parameter but, when κ increases, the AFE of the vMF distribution increases faster. Additionally, even if the AFE are close for small values of κ, the distributions of the

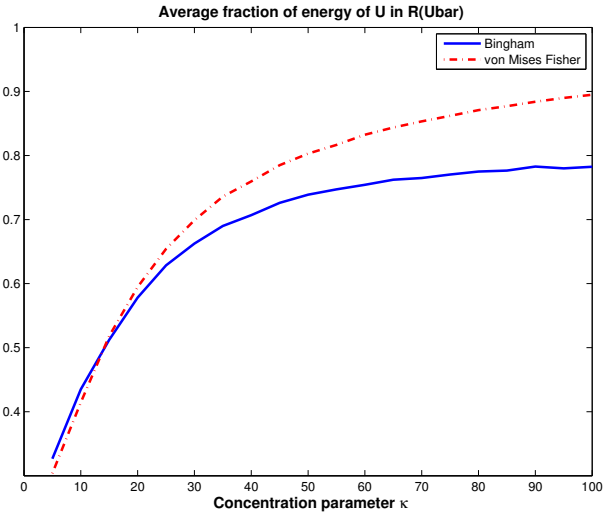

Figure 1: Average fraction of energy of $\boldsymbol{U}$ in $\mathcal{R}(\bar{\boldsymbol{U}})$ versus κ. $N = 20$, $p = 5$.

angles between $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$ exhibit some differences, as shown in Figures 2 and 3 which display the probability density functions of these angles for $\kappa = 20$.

3.2 Linear model

In order to illustrate how the previous theory can be used in practice, we first consider a simple example, namely a linear Gaussian model (conditioned on $\boldsymbol{U}$), i.e., we assume that the data follows the model $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{S} + \boldsymbol{N}$ where the columns of $\boldsymbol{N}$ are independent and identically distributed (i.i.d.) Gaussian vectors with zero-mean and (known) covariance matrix $\sigma_n^2 \boldsymbol{I}$. We assume that no knowledge about $\boldsymbol{S}$ is available and hence its prior distribution is set to $\pi(\boldsymbol{S}) \propto 1$. Therefore, conditioned on $\boldsymbol{U}$ we have

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{U}) &= \int p(\boldsymbol{Y}|\boldsymbol{U}, \boldsymbol{S})\, \pi(\boldsymbol{S})\, d\boldsymbol{S} \\ &\propto \int \text{etr}\left\{-\frac{1}{2\sigma_n^2}(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{S})^T (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{S})\right\} d\boldsymbol{S} \\ &\propto \text{etr}\left\{-\frac{1}{2\sigma_n^2}\boldsymbol{Y}^T\boldsymbol{Y} + \frac{1}{2\sigma_n^2}\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Y}\right\}. \end{aligned} \tag{12}$$

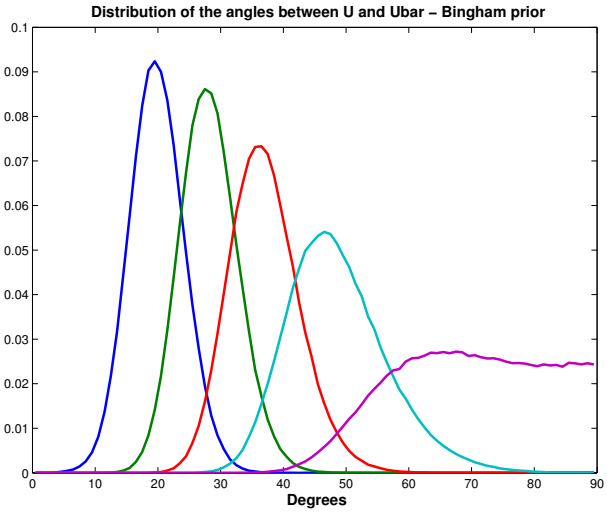

Figure 2: Distribution of the angles between $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$ for a Bingham distribution. $N = 20$, $p = 5$ and $\kappa = 20$.

The above distribution, along with the prior distribution $\pi(\boldsymbol{U})$, is now used to derive the MMSD estimator. We successively investigate the case of a Bingham prior and that of a vMF prior.

Proposition 1. *When $\boldsymbol{U}$ is assigned the Bingham prior distribution, the MMSD estimator is obtained in* ***closed-form*** *as*

$$\hat{\boldsymbol{U}}_{mmsd\text{-}LM\text{-}B} = \mathcal{P}_p\left\{\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T\right\}. \tag{13}$$

The proof of this proposition is given in Appendix A. Therefore, in this case, the MMSD estimator has a very simple form. It consists of the principal subspace of a (weighted) combination of the a priori projection matrix $\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T$ and the information brought by the data through $\boldsymbol{Y}\boldsymbol{Y}^T$. Observe that, in this particular case of a Bingham posterior, the MMSD estimator coincides with the maximum a posteriori (MAP) estimator.

Let us now consider the case where the prior distribution of $\boldsymbol{U}$ is vMF, and contrast it with the previous example. Using (12) along with along with (10), it follows that the posterior distribution now writes

$$p(\boldsymbol{U}|\boldsymbol{Y}) \propto \text{etr}\left\{\kappa\boldsymbol{U}^T\bar{\boldsymbol{U}} + \frac{1}{2\sigma_n^2}\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{U}\right\} \tag{14}$$

which is referred to as the Bingham-von-Mises-Fisher (BMF) distribution with parameter matrices $\boldsymbol{Y}\boldsymbol{Y}^T$, $\frac{1}{2\sigma_n^2}\boldsymbol{I}$ and $\kappa\bar{\boldsymbol{U}}$ respectively[1]. Although this distribution is known [27], to our knowledge, there does not exist any analytic expression for the integral in (4) when $\boldsymbol{U}|\boldsymbol{Y}$ has the BMF distribution (14). Therefore, the MMSD estimator cannot be computed in closed-form. In such situation, it is very common to implement a Markov chain Monte Carlo (MCMC) method for sampling according to the posterior distribution of interest. There are many MCMC algorithms that could be used for that purpose (the reader is invited to consult [32,33] for more details). However, when the full conditional distributions of the target posterior distribution can be sampled, the very popular Gibbs sampler is generally adopted for simplicity. An efficient Gibbs sampling scheme to generate random unitary matrices drawn from a BMF $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$

[1]The matrix $\boldsymbol{X}$ is said to have a BMF $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ distribution -where $\boldsymbol{A}$ is an $N \times N$ symmetric matrix, $\boldsymbol{B}$ is a $p \times p$ diagonal matrix and $\boldsymbol{C}$ is an $N \times p$ matrix- if $p(\boldsymbol{X}) \propto \text{etr}\left\{\boldsymbol{C}^T\boldsymbol{X} + \boldsymbol{B}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\right\}$.

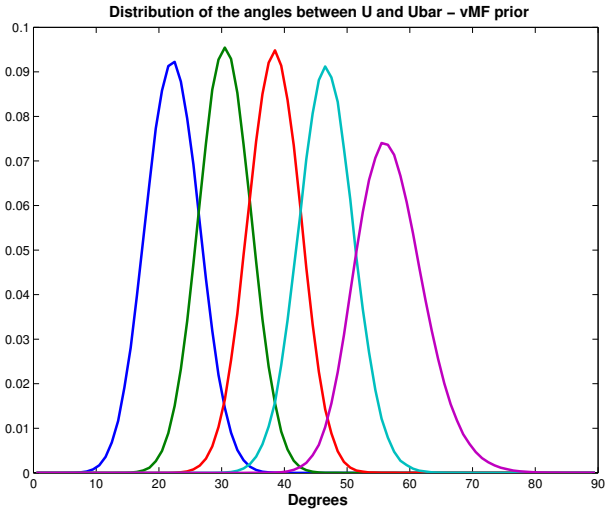

Figure 3: Distribution of the angles between $\mathcal{R}(\boldsymbol{U})$ and $\mathcal{R}(\bar{\boldsymbol{U}})$ for a von Mises Fisher distribution. $N = 20$, $p = 5$ and $\kappa = 20$.

distribution with arbitrary full-rank matrix $\boldsymbol{A}$ was proposed in [34]. It amounts to sampling successively each column of $\boldsymbol{U}$ by generating a random unit norm vector drawn from a (vector) BMF distribution. In our case, $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^T$ whose rank is $\min(K, N)$ and hence $\boldsymbol{A}$ is rank-deficient whenever $K < N$, a case of most interest to us. Note also that to generate matrices $\boldsymbol{U}$ drawn from the Bingham distribution in (9), we need to consider $\boldsymbol{A} = \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T$ which has rank $p < N$. Therefore, the scheme of [34] needs to be adapted in order to generate random matrices drawn from (14): details on how this can be achieved can be found in Appendix B. Once these matrices asymptotically distributed according to (14) are obtained, (4) can be approximated as

$$\hat{\boldsymbol{U}}_{\text{mmsd-LM-vMF}} \simeq \mathcal{P}_p \left\{ \frac{1}{N_r} \sum_{n=N_{\text{bi}}+1}^{N_{\text{bi}}+N_r} \boldsymbol{U}^{(n)} \boldsymbol{U}^{(n)H} \right\}. \tag{15}$$

In (15), N_{bi} is the number of burn-in samples and N_r is the number of samples used to approximate the estimator.

Remark 3. Interestingly enough, the above estimator in (15) is the so-called induced arithmetic mean (IAM) [35] of the set of unitary matrices $\boldsymbol{U}^{(n)}$, $n = N_{\text{bi}} + 1, cdots, N_{\text{bi}} + N_r$. It differs from the Karcher mean of the set $\boldsymbol{U}^{(n)}$, $n = N_{\text{bi}} + 1, \cdots, N_{\text{bi}} + N_r$, which truly minimizes the sum of the distances to all $\boldsymbol{U}^{(n)}$. However, the Karcher mean may not exist and requires iterative schemes to be computed [36] while the IAM is straightforward to compute.

Remark 4. In the particular case where $\boldsymbol{U}$ has a Bingham prior distribution, the MAP estimator of $\boldsymbol{U}$ and its MMSD estimator are equal. This is no longer true when $\boldsymbol{U}$ has a vMF prior distribution, and hence a BMF posterior distribution. The mode of the latter is not known in closed-form either. However, it can be approximated by selecting, among the matrices generated by the Gibbs sampler, the matrix which results in the largest value of the posterior distribution.

3.3 Covariance matrix model

We now consider a more complicated case where $\boldsymbol{Y}$, conditioned on $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$, is Gaussian distributed with zero-mean and covariance matrix

$$\boldsymbol{R} = \mathsf{E}\left\{\boldsymbol{Y}\boldsymbol{Y}^T\right\} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T + \sigma_n^2 \boldsymbol{I} \tag{16}$$

where $\boldsymbol{U}$ is an orthonormal basis for the signal subspace, $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues and σ_n^2 stands for the white noise power which is assumed to be known here. As it will be more convenient and more intuitively appealing, we re-parametrize the covariance matrix as follows. The inverse of $\boldsymbol{R}$ can be written as

$$\begin{aligned}\boldsymbol{R}^{-1} &= \boldsymbol{U}\left[\left(\boldsymbol{\Lambda}+\sigma_n^2\boldsymbol{I}\right)^{-1}-\sigma_n^{-2}\boldsymbol{I}\right]\boldsymbol{U}^T+\sigma_n^{-2}\boldsymbol{I}\\ &= \sigma_n^{-2}\boldsymbol{I}-\sigma_n^{-2}\boldsymbol{U}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}+\sigma_n^2\boldsymbol{I}\right)^{-1}\boldsymbol{U}^T\\ &\triangleq \nu\boldsymbol{I}-\nu\boldsymbol{U}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\boldsymbol{U}^T \end{aligned} \tag{17}$$

where $\nu \triangleq \sigma_n^{-2}$, $\boldsymbol{\Gamma} \triangleq \mathrm{diag}\left(\boldsymbol{\gamma}\right)$ with $\boldsymbol{\gamma} = \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_p\end{bmatrix}^T$ and

$$0 < \gamma_k \triangleq \frac{\sigma_n^2}{\sigma_n^2+\lambda_k} < 1. \tag{18}$$

The idea is to parametrize the problem in terms of $\boldsymbol{U}$ and $\boldsymbol{\Gamma}$ rather than $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$. The interest of this transformation is twofold. First, it enables one to express all eigenvalues with respect to the white noise level. Indeed, one has $\boldsymbol{R} = \nu^{-1}\boldsymbol{U}_\perp\boldsymbol{U}_\perp^T + \nu^{-1}\boldsymbol{U}\boldsymbol{\Gamma}^{-1}\boldsymbol{U}^T$ where $\boldsymbol{U}_\perp$ is an orthonormal basis for $\mathcal{R}\left(\boldsymbol{U}\right)^\perp$ and hence the γ_ks are representative of the scaling between the "signal" eigenvalues and the noise eigenvalues. In fact, they carry information about the signal-to-noise ratio since $\gamma_k = \left(1+\frac{\lambda_k}{\sigma^2}\right)^{-1}$ and $\frac{\lambda_k}{\sigma^2}$ represents the SNR of the k-th signal component. Second, this new parametrization will facilitate derivation of the conditional distributions required for the Gibbs sampler.

Since $\boldsymbol{Y}$ conditioned on $\boldsymbol{U}$ and $\boldsymbol{\gamma}$ is Gaussian, it follows that

$$p\left(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\gamma}\right) = \left(2\pi\right)^{-NK/2}\left|\boldsymbol{R}\right|^{-K/2}\mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{Y}^T\boldsymbol{R}^{-1}\boldsymbol{Y}\right\}. \tag{19}$$

From $\boldsymbol{R}^{-1} = \nu\boldsymbol{U}_\perp\boldsymbol{U}_\perp^T + \nu\boldsymbol{U}\boldsymbol{\Gamma}\boldsymbol{U}^T$, it ensues that $\left|\boldsymbol{R}^{-1}\right| = \nu^N\left|\boldsymbol{\Gamma}\right|$ and hence

$$p\left(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\gamma}\right) \propto \left|\boldsymbol{\Gamma}\right|^{K/2}\mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{Y}^T\left[\nu\boldsymbol{I}-\nu\boldsymbol{U}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\boldsymbol{U}^T\right]\boldsymbol{Y}\right\}. \tag{20}$$

Let us now consider the prior distributions for $\boldsymbol{U}$ and $\boldsymbol{\gamma}$. We will consider either a Bingham or vMF distribution for $\boldsymbol{U}$. As for $\boldsymbol{\gamma}$, we assume that γ_k are a priori independent random variables uniformly distributed in the interval $[\gamma_-,\gamma_+]$, i.e.,

$$\pi(\boldsymbol{\gamma}) = \prod_{k=1}^{p}\left(\gamma_+-\gamma_-\right)^{-1}\mathbb{I}_{[\gamma_-,\gamma_+]}(\gamma_k). \tag{21}$$

The value of γ_+ [respectively γ_-] can be set to 1 [respectively 0] if a non-informative prior is desired. Otherwise, if some information is available about the SNR, γ_- and γ_+ can be chosen so as to reflect this knowledge since $\gamma_k = \left(1+SNR_k\right)^{-1}$: γ_+ [resp. γ_-] rules the lowest [resp. highest] value of the SNR, say SNR_- [resp. SNR_+].

As explained in Appendix C, marginalization of $p\left(\boldsymbol{U},\boldsymbol{\gamma}|\boldsymbol{Y}\right)$ with respect to $\boldsymbol{\gamma}$ leads to intractable distributions $p(\boldsymbol{U}|\boldsymbol{Y})$. Therefore, in order to implement the MMSD estimator, we propose to draw samples from the the joint posterior distribution of $\boldsymbol{U}$ and $\boldsymbol{\gamma}$, and then average them to estimate $\boldsymbol{U}$, similarly to what is done in (15). In Appendix C, we provide the details for generating samples according to $p\left(\boldsymbol{U},\boldsymbol{\gamma}|\boldsymbol{Y}\right)$, which is given by

Bingham prior

$$\begin{aligned} p\left(\boldsymbol{U},\boldsymbol{\gamma}|\boldsymbol{Y}\right) &\propto p\left(\boldsymbol{Y}|\boldsymbol{U},\boldsymbol{\gamma}\right)\pi\left(\boldsymbol{U}\right)\pi\left(\boldsymbol{\gamma}\right) \\ &\propto \left|\boldsymbol{\Gamma}\right|^{K/2}\left(\prod_{k=1}^{p}\mathbb{I}_{[\gamma_-,\gamma_+]}(\gamma_k)\right) \\ &\times \operatorname{etr}\left\{\kappa\boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U}+\frac{\nu}{2}\boldsymbol{Y}^T\boldsymbol{U}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\boldsymbol{U}^T\boldsymbol{Y}\right\}. \end{aligned} \tag{22}$$

vMF prior

$$\begin{aligned} p\left(\boldsymbol{U},\boldsymbol{\gamma}|\boldsymbol{Y}\right) &\propto \left|\boldsymbol{\Gamma}\right|^{K/2}\left(\prod_{k=1}^{p}\mathbb{I}_{[\gamma_-,\gamma_+]}(\gamma_k)\right) \\ &\times \operatorname{etr}\left\{\kappa\boldsymbol{U}^T\bar{\boldsymbol{U}}+\frac{\nu}{2}\boldsymbol{Y}^T\boldsymbol{U}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\boldsymbol{U}^T\boldsymbol{Y}\right\}. \end{aligned} \tag{23}$$

4 Simulations

In this section we illustrate the performance of the approach developed above through Monte Carlo simulations. In all simulations $N = 20$, $p = 5$ and $\kappa = 20$. The matrix $\boldsymbol{S}$ is generated from a Gaussian distribution with zero-mean and covariance matrix $\sigma_s^2\boldsymbol{I}$ and the signal-to-noise ratio is defined as $SNR = 10\log_{10}\left(\sigma_s^2/\sigma_n^2\right)$. The matrix $\boldsymbol{U}$ is generated from the Bingham distribution (9) or the vMF distribution (10) and, for the sake of simplicity, $\bar{\boldsymbol{U}} = \begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{0}\end{bmatrix}^T$. The number of burn-in iterations in the Gibbs sampler is set to $N_{\text{bi}} = 10$ and $N_r = 1000$. The MMSD estimator (4) is compared with the MAP estimator, the MMSE estimator, the usual SVD-based estimator and the estimator $\hat{\boldsymbol{U}} = \bar{\boldsymbol{U}}$ that discards the available data and use only the a priori knowledge. The latter is referred to as "Ubar" in the figures. The estimators are evaluated in terms of the fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}\left(\boldsymbol{U}\right)$, i.e., $\text{AFE}\left(\hat{\boldsymbol{U}},\boldsymbol{U}\right)$.

4.1 Linear model

We begin with the linear model. Figures 4 to 7 investigate the influence of K and SNR onto the performance of the estimators. Figures 4 and 5 concern the Bingham prior while the vMF prior has been used to obtain Figures 6 and 7. From inspection of these figures, the following conclusions can be drawn:

- the MMSD estimator performs better than the estimator $\hat{\boldsymbol{U}} = \bar{\boldsymbol{U}}$, even at low SNR. The improvement is all the more pronounced that K is large. Therefore, the MMSD estimator makes a sound use of the data to improve accuracy compared to using the prior knowledge only.
- the MMSD estimator performs better than the SVD, especially at low SNR. Moreover, and this is a distinctive feature of this Bayesian approach, it enables one to estimate the subspace even when the number of snapshots K is less than the size of the subspace p.
- for a Bingham prior, the MMSE performs very poorly since the posterior distribution of $\boldsymbol{U}$ conditioned on $\boldsymbol{Y}$ depends on $\boldsymbol{U}\boldsymbol{U}^T$ only. Hence, averaging the matrix $\boldsymbol{U}$ itself does not make sense, see our remark 1. In contrast, when $\boldsymbol{U}$ has a vMF prior, the posterior depends on both $\boldsymbol{U}$ and $\boldsymbol{U}\boldsymbol{U}^T$: in this case, the MMSE performs well and is close to the MMSD. Note however that the vMF prior is more restrictive than the Bingham prior.
- the MMSD estimator also outperforms the MAP estimator.

As a conclusion, the MMSD estimator performs better than most other estimators in the large majority of cases.

4.2 Covariance matrix model

We now conduct simulations with the covariance matrix model. The simulation parameters are essentially the same as in the previous section, except for the SNR. More precisely, the random variables γ_k are drawn from the uniform distribution in (21) where γ_- and γ_+ are selected such that $SNR_- = 5$dB and $SNR_+ = 10$dB. The results are shown in Fig. 8 for the Bingham prior and Fig. 9 for the vMF prior. They corroborate the previous observations made on the linear model, viz that the MMSD estimator offers the best performance over all methods.

5 Application to hyperspectral imagery

In this section, we show how the proposed subspace estimation procedure can be efficiently used for an application to multi-band image analysis. For several decades, hyperspectral imagery has received considerable attention because of its great interest for various purposes: agriculture monitoring, mineral mapping, military concerns, etc. One of the crucial issue when analyzing such image is the spectral unmixing which aims to decompose an observed pixel $\boldsymbol{y}_\ell$ into a collection of $R = p + 1$ reference signatures, $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_R$ (called *endmembers*) and to retrieve the respective proportions of these signatures (or *abundances*) $a_{1,\ell}, \ldots, a_{R,\ell}$ in this pixel [37]. To describe the physical process that links the endmembers and their abundances to the measurements, the most widely admitted mixing model is linear

$$\boldsymbol{y}_\ell = \sum_{r=1}^{R} a_{r,\ell} \boldsymbol{m}_r \tag{24}$$

where $\boldsymbol{y}_\ell \in \mathbb{R}^N$ is the pixel spectrum measured in N spectral bands, $\boldsymbol{m}_r \in \mathbb{R}^N$ $(r = 1, \ldots, R)$ are the R endmember spectra and $a_{r,\ell}$ $(r = 1, \ldots, R)$ are their corresponding abundances. Due to obvious physical considerations, the abundances obey two kinds of constraints. Since they represent proportions, they must satisfy the following positivity and additivity constraints

$$\begin{cases} a_{r,\ell} \geq 0, \quad r = 1, \ldots, R, \\ \sum_{r=1}^{R} a_{r,\ell} = 1. \end{cases} \tag{25}$$

Let now consider L pixels $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$ of an hyperspectral image induced by the linear mixing model (LMM) in (24) with the abundance constraints (25). It is clear that the dataset formed by these L pixels lies in a lower-dimensional subspace $\mathcal{U} \subset \mathbb{R}^p$. More precisely, in this subspace $\mathcal{U}$, the dataset belongs to a simplex whose vertices are the endmembers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_R$ to be recovered. Most of the unmixing strategies developed in the hyperspectral imagery literature are based on this underlying geometrical formulation of the LMM. Indeed, the estimation of the endmembers is generally conducted in the lower-dimensional space $\mathcal{U}$, previously identified by a standard dimension reduction technique such as the principal component analysis (PCA) [37]. However, it is well known that the model linearity is a simplifying assumption and does not hold anymore in several contexts, circumventing the standard unmixing algorithms. Specifically, non- inearities are known to occur for scenes including mixtures of minerals or vegetation. As a consequence, evaluating the suitability of the LMM assumption for a given hyperspectral image is a capital question that can be conveniently addressed by the approach introduced above.

5.1 Synthetic data

First, we investigate the estimation of the subspace $\mathcal{U}$ when the image pixels are non-linear functions of the abundances. For this purpose, a 50×50 synthetic hyperspectral image is generated following a recently introduced non-linear model referred to as generalized bilinear

model (GBM). As indicated in [38], the GBM is notably well adapted to describe non-linearities due to multipath effects. It assumes that the observed pixel spectrum $\boldsymbol{y}_\ell$ can be written

$$\boldsymbol{y}_\ell = \sum_{r=1}^{R} a_{r,\ell}\boldsymbol{m}_r + \sum_{i=1}^{R-1}\sum_{j=i+1}^{R} \gamma_{i,j,\ell} a_{i,\ell} a_{j,\ell} \boldsymbol{m}_i \odot \boldsymbol{m}_j \tag{26}$$

where $\odot$ stands for the Hadamard (termwise) product and the abundances $a_{r,\ell}$ ($r = 1, \ldots, R$) satisfy the constraints in (25). In (26), the parameters $\gamma_{i,j,\ell}$ (which belong to $[0, 1]$) characterize the importance of non-linear interactions between the endmembers $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ in the ℓ-th pixel. In particular, when $\gamma_{i,j,\ell} = 0$ ($\forall i, j$), the GBM reduces to the standard LMM (24). Moreover, when $\gamma_{i,j,\ell} = 1$ ($\forall i, j$), the GBM leads to the non-linear model introduced by Fan *et al.* in [39]. In this simulation, the synthetic image has been generated using the GBM with $R = 3$ endmember signatures extracted from a spectral library. The corresponding abundances have been uniformly drawn in the set defined by the constraints (25). We have assumed that there is no interaction between endmembers $\boldsymbol{m}_1$ and $\boldsymbol{m}_3$, and between endmembers $\boldsymbol{m}_2$ and $\boldsymbol{m}_3$ resulting in $\gamma_{1,3,\ell} = \gamma_{2,3,ell} = 0$, $\forall \ell$. Moreover, the interactions between endmembers $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are defined by the map of coefficients $\gamma_{1,2,\ell}$ displayed in Fig. 10 (top, left panel) where a black (resp. white) pixel represents the lowest (resp. highest) degree of non-linearity. As can be seen in this figure, 75% of the pixels (located in the bottom and upper right squares of the image) are mixed according to the LMM resulting in $\gamma_{1,2,\ell} = 0$. The 25% remaining image pixels (located in the upper left square of the image) are mixed according to the GBM with nonlinearity coefficients $\gamma_{1,2,\ell}$ radially increasing from 0 to 1 ($\gamma_{1,2,\ell} = 0$ in the image center and $\gamma_{1,2,\ell} = 1$ in the upper left corner of the image). Note that this image contains a majority of pixels that are mixed linearly and belong to a common subspace of $\mathbb{R}^2$. Conversely, the non-linearly mixed pixels do not belong to this subspace[2]. We propose here to estimate the local subspace $\mathcal{U}_\ell$ where a given image pixel $\boldsymbol{y}_\ell$ and its nearest spectral neighbors $\mathcal{V}_\ell^{(K-1)}$ live ($\mathcal{V}_\ell^{(K-1)}$ denotes the set of the $(K-1)$-nearest neighbors of $\boldsymbol{y}_\ell$).

Assuming as a first approximation that all the image pixels are linearly mixed, all these pixels are approximately contained in a common 2-dimensional subspace $\bar{\mathcal{U}}$ that can be determined by performing a PCA of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$ (see [40] for more details). The corresponding principal vectors spanning $\bar{\mathcal{U}}$ are gathered in a matrix $\bar{\boldsymbol{U}}$. This matrix $\bar{\boldsymbol{U}}$ is used as *a priori* knowledge regarding the 2-dimensional subspace containing $\left\{\boldsymbol{y}_\ell, \mathcal{V}_\ell^{(K-1)}\right\}_{\ell=1,\ldots,L}$. However, this crude estimation can be refined by the Bayesian estimation strategy developed in the previous sections. More precisely, for each pixel $\boldsymbol{y}_\ell$, we compute the MMSD estimator of the $N \times p$ matrix $\boldsymbol{U}_\ell$, whose columns are supposed to span the subspace $\mathcal{U}_\ell$ containing $\boldsymbol{y}_\ell$ and its $K-1$-nearest neighbors $\mathcal{V}_\ell^{(K-1)}$. The Bayesian estimator $\hat{\boldsymbol{U}}_\ell$ is computed from its closed-form expression (13), i.e., using the Bingham prior where $\bar{\boldsymbol{U}}$ has been introduced above. Then, for each pixel, we evaluate the distance between the two projection matrices $\hat{\boldsymbol{U}}_\ell \hat{\boldsymbol{U}}_\ell^T$ and $\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T$ onto the subspaces $\hat{\mathcal{U}}_\ell = \mathcal{R}\left(\hat{\boldsymbol{U}}_\ell\right)$ and $\bar{\mathcal{U}} = \mathcal{R}\left(\bar{\boldsymbol{U}}\right)$, respectively. As stated in Section 2, the natural distance between these two projection matrices is given by $d^2\left(\hat{\boldsymbol{U}}_\ell, \bar{\boldsymbol{U}}\right) = 2\left(p - \mathrm{Tr}\left\{\hat{\boldsymbol{U}}_\ell^T \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T \hat{\boldsymbol{U}}_\ell\right\}\right)$. The resulting distance maps are depicted in Fig. 10 (bottom panels) for 2 non-zero values of $\eta \triangleq 2\sigma_n^2\kappa$ (as it can be noticed in (13), this hyperparameter η balances the quantity of *a priori* knowledge $\bar{\boldsymbol{U}}$ included in the estimation with respect to the information brought by the data). For comparison purpose, the subspace $\hat{\mathcal{U}}_\ell$ has been also estimated by a crude SVD of $\left\{\boldsymbol{y}_\ell, \mathcal{V}_\ell^{(K-1)}\right\}$ (top right panel). In this case, $\hat{\boldsymbol{U}}_\ell$ simply reduces to the associated principal singular vectors and can be considered as the MMSD estimator of $\boldsymbol{U}_\ell$ obtained for $\eta = 0$.

[2] Assuming there is a majority of image pixels that are mixed linearly is a reasonable assumption for most hyperspectral images.

These figures show that, for the 75% of the pixels generated using the LMM (bottom and right parts of the image), the subspace $\bar{\mathcal{U}}$ estimated by an SVD of the whole dataset $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$ is very close to the hyperplanes $\hat{\mathcal{U}}_\ell$ locally estimated from $\left\{\boldsymbol{y}_\ell, \mathcal{V}_\ell^{(K-1)}\right\}$ through the proposed approach (for any value of η). Regarding the remaining 25% pixels resulting from the GBM (top left part of the image), the following comments can be made. When a crude SVD of $\left\{\boldsymbol{y}_\ell, \mathcal{V}_\ell^{(K-1)}\right\}$ is conducted, i.e., when no prior knowledge is taken into account to compute the MMSD ($\eta = 0$, top right panel), the distance between the locally estimated subspace $\hat{\mathcal{U}}_\ell$ and the *a priori* assumed hyperplane $\bar{\mathcal{U}}$ does not reflect the non-linearities contained in the image. Conversely, when this crude SVD is regularized by incorporating prior knowledge with $\eta = 0.5$ and $\eta = 50$ (bottom left and right panels, respectively), leading to the MMSD estimator, the larger the degree of non-linearity, the larger the distance between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_\ell$. To summarize, evaluating the distance between the MMSD estimator $\hat{\boldsymbol{U}}_\ell$ and the *a priori* given matrix $\bar{\boldsymbol{U}}$ allows the degree of non-linearity to be quantified. This interesting property is exploited on a real hyperspectral image in the following section.

5.2 Real data

The real hyperspectral image considered in this section has been acquired in 1997 over Moffett Field, CA, by the NASA spectro-imager AVIRIS. This image, depicted with composite true colors in Fig. 11 (top, left panel), has been minutely studied in [40] assuming a linear mixing model. The scene consists of a large part of a lake (black pixels, top) and a coastal area (bottom) composed of soil (brown pixels) and vegetation (green pixels), leading to $R = 3$ endmembers whose spectra and abundance maps can be found in [40]. A simple estimation of a lower-dimensional space $\bar{\mathcal{U}}$ where the pixels live can be conducted through a direct SVD of the whole dataset, providing the *a priori* matrix $\bar{\boldsymbol{U}}$. As in the previous section, this crude estimation can be refined by computing locally the MMSD estimators $\hat{\boldsymbol{U}}_\ell$ spanning the subspaces $\hat{\mathcal{U}}_\ell$ (bottom panels). These estimators have been also computed with $\eta = 0$, corresponding to an SVD of $\left\{\boldsymbol{y}_\ell, \mathcal{V}_\ell^{(K-1)}\right\}$ (top, right figure). The distances between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_\ell$ have been reported in the maps of Fig. 11. Again, for $\eta = 0$ (top, right panel), a simple local SVD is unable to locate possible non-linearities in the scene. However, for two[3] non-zero values $\eta = 0.5$ and $\eta = 50$ (bottom left and right panels, respectively), the distances between the *a priori* recovered subspace $\bar{\mathcal{U}}$ and the MMSD-based subspace $\hat{\mathcal{U}}_\ell$ clearly indicate that some non-linear effects occur in specific parts of the image, especially in the lake shore. Note that the non-linearities identified by the proposed algorithm are very similar to the ones highlighted in [38] where the unmixing procedure was conducted by using the GBM defined in (26). This shows the accuracy of the proposed MMSD estimator to localize the non-linearities occurring in the scene, which is interesting for the analysis of hyperspectral images.

6 Conclusions

This paper considered the problem of estimating a subspace using some available a priori information. Towards this end, a Bayesian framework was advocated, where the subspace $\boldsymbol{U}$ is assumed to be drawn from an appropriate prior distribution. However, since we operate in a Grassmann manifold, the conventional MMSE approach is questionable as it amounts to minimizing a distance which is not the most meaningful on the Grassmann manifold. Consequently, we revisited the MMSE approach and proposed, as an alternative, to minimize a natural distance

[3] Additional results obtained with other values of η are available online at `http://dobigeon.perso.enseeiht.fr/app_MMSD.html`.

on the Grassmann manifold. A general framework was formulated resulting in a novel estimator which entails computing the principal eigenvectors of the posterior mean of $\boldsymbol{U}\boldsymbol{U}^T$. The theory was exemplified on a few simple examples, where the MMSD estimator can either be obtained in closed- orm or requires resorting to an MCMC simulation method. The new approach enables one to combine efficiently the prior knowledge and the data information, resulting in a method that performs well at low SNR or with very small sample support. A successful application to the analysis of non-linearities contained in hyperspectral images was also presented.

A Proof of Proposition 1

In this appendix, we derive the MMSD estimator for the linear model when $\boldsymbol{U}$ follows a Bingham prior. In this case, the posterior distribution of $\boldsymbol{U}$, conditioned on $\boldsymbol{Y}$ is given by

$$p\left(\boldsymbol{U}|\boldsymbol{Y}\right) \propto \operatorname{etr}\left\{\boldsymbol{U}^T\left[\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T\right]\boldsymbol{U}\right\} \tag{27}$$

which is recognized as a Bingham distribution with parameter matrix $\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T$, i.e., $\boldsymbol{U}|\boldsymbol{Y} \sim \mathrm{B}\left(\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T\right)$. Therefore, in order to derive the MMSD estimator, we need to derive the eigenvectors of $\int \boldsymbol{U}\boldsymbol{U}^T p\left(\boldsymbol{U}|\boldsymbol{Y}\right) d\boldsymbol{U}$ when $p\left(\boldsymbol{U}|\boldsymbol{Y}\right)$ is a Bingham distribution. Towards this end, we make use of the following proposition.

Proposition 2. *Let $\boldsymbol{U} \in \mathbb{R}^{N\times p}$ be an orthogonal matrix -$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$- drawn from a Bingham distribution with parameter matrix $\boldsymbol{A}$*

$$p_{\mathrm{B}}(\boldsymbol{U}) = \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\operatorname{etr}\left\{\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}\right\} \tag{28}$$

with $\kappa_{\mathrm{B}}(\boldsymbol{A}) = \ln {}_1F_1\left(\frac{1}{2}p, \frac{1}{2}N; \boldsymbol{A}\right)$. Let $\boldsymbol{A} = \boldsymbol{U}_a\boldsymbol{\Lambda}_a\boldsymbol{U}_a^T$ denote the eigenvalue decomposition of $\boldsymbol{A}$ where the eigenvalues are ordered in descending order. Let us define $\boldsymbol{M} = \int \boldsymbol{U}\boldsymbol{U}^T p_{\mathrm{B}}(\boldsymbol{U}) d\boldsymbol{U}$. Then the eigenvalue decomposition of $\boldsymbol{M}$ writes

$$\boldsymbol{M} = \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\boldsymbol{U}_a\boldsymbol{\Gamma}\boldsymbol{U}_a^T$$

with $\boldsymbol{\Gamma} = \frac{\partial \exp\{\kappa_{\mathrm{B}}(\boldsymbol{A})\}}{\partial \boldsymbol{\Lambda}_a}$ and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_N$ where $\gamma_n = \boldsymbol{\Gamma}(n,n)$.

Proof. For notational convenience, let us work with the projection matrix $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^T$ whose distribution on the Grassmann manifold is [27]

$$p(\boldsymbol{P}) = \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\operatorname{etr}\left\{\boldsymbol{P}\boldsymbol{A}\right\}. \tag{29}$$

We have then that

$$\begin{aligned}
\boldsymbol{M} &= \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\int \boldsymbol{P}\operatorname{etr}\left\{\boldsymbol{P}\boldsymbol{U}_a\boldsymbol{\Lambda}_a\boldsymbol{U}_a^T\right\} d\boldsymbol{P} \\
&= \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\boldsymbol{U}_a\left[\int \boldsymbol{U}_a^T\boldsymbol{P}\boldsymbol{U}_a\operatorname{etr}\left\{\boldsymbol{U}_a^T\boldsymbol{P}\boldsymbol{U}_a\boldsymbol{\Lambda}_a\right\} d\boldsymbol{P}\right]\boldsymbol{U}_a^T \\
&= \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\boldsymbol{U}_a\left[\int \boldsymbol{P}\operatorname{etr}\left\{\boldsymbol{P}\boldsymbol{\Lambda}_a\right\} d\boldsymbol{P}\right]\boldsymbol{U}_a^T \\
&= \exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\boldsymbol{U}_a\boldsymbol{\Gamma}\boldsymbol{U}_a^T.
\end{aligned}$$

Moreover $\boldsymbol{\Gamma}$ is diagonal since, for any orthogonal diagonal matrix $\boldsymbol{D}$,

$$\begin{aligned}
\boldsymbol{\Gamma}\boldsymbol{D} &= \int \boldsymbol{P}\boldsymbol{D}\operatorname{etr}\left\{\boldsymbol{P}\boldsymbol{\Lambda}_a\right\} d\boldsymbol{P} \\
&= \boldsymbol{D}\left[\int \boldsymbol{D}^T\boldsymbol{P}\boldsymbol{D}\operatorname{etr}\left\{\boldsymbol{D}^T\boldsymbol{P}\boldsymbol{D}\boldsymbol{D}^T\boldsymbol{\Lambda}_a\boldsymbol{D}\right\} d\boldsymbol{P}\right] \\
&= \boldsymbol{D}\int \boldsymbol{P}\operatorname{etr}\left\{\boldsymbol{P}\boldsymbol{\Lambda}_a\right\} d\boldsymbol{P} \\
&= \boldsymbol{D}\boldsymbol{\Gamma}
\end{aligned}$$

where, to obtain the third line, we made use of the fact that $\boldsymbol{D}^T\boldsymbol{\Lambda}_a\boldsymbol{D} = \boldsymbol{\Lambda}_a$. It follows that the eigenvectors of $\boldsymbol{M}$ and $\boldsymbol{A}$ coincide, and that the eigenvalues of $\boldsymbol{M}$ are $\exp\left\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\right\}\gamma_n$, for

$n = 1, \cdots, N$. Moreover, it is known that $\exp\{-\kappa_{\mathrm{B}}(\boldsymbol{A})\} = \exp\{-\kappa_{\mathrm{B}}(\boldsymbol{\Lambda}_a)\}$ and, from (29), one has

$$\exp\{\kappa_{\mathrm{B}}(\boldsymbol{\Lambda}_a)\} = \int \mathrm{etr}\{\boldsymbol{P}\boldsymbol{\Lambda}_a\}\, d\boldsymbol{P}.$$

Differentiating the latter equation with respect to $\lambda_a(k)$ and denoting $p_n = \boldsymbol{P}(n,n)$, one obtains

$$\begin{aligned}\frac{\partial \exp\{\kappa_{\mathrm{B}}(\boldsymbol{\Lambda}_a)\}}{\partial \lambda_a(k)} &= \frac{\partial}{\partial \lambda_a(k)} \int \exp\left\{\sum_{n=1}^{N} \lambda_a(n) p_n\right\} d\boldsymbol{P} \\ &= \int p_k \mathrm{etr}\{\boldsymbol{P}\boldsymbol{\Lambda}_a\}\, d\boldsymbol{P} \\ &= \gamma_k.\end{aligned}$$

The previous equation enables one to relate the eigenvalues of $\boldsymbol{A}$ and those of $\boldsymbol{M}$. It remains to prove that $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_N$. Towards this end, we make use of a very general theorem due to Letac [41], which is briefly outlined below. Let

$$P(\mu, \boldsymbol{A})(d\boldsymbol{X}) = \exp\{\kappa_\mu(\boldsymbol{A})\}\, \mathrm{etr}\left\{\boldsymbol{X}^T\boldsymbol{A}\right\} \mu(d\boldsymbol{X})$$

be a probability associated with a unitarily invariant measure μ on the set of $N \times N$ symmetric matrices. Consider the case of a diagonal matrix $\boldsymbol{A} = \mathrm{diag}\,(a_1, a_2, \cdots, a_N)$ with $a_1 \geq a_2 \geq \cdots \geq a_N$. Then [41] proves that $\boldsymbol{M} = \int \boldsymbol{X} P(\mu, \boldsymbol{A})(d\boldsymbol{X})$ is also diagonal, and moreover if $\boldsymbol{M} = \mathrm{diag}\,(m_1, m_2, \cdots, m_N)$ then $m_1 \geq m_2 \geq \cdots \geq m_N$. Use of this theorem completes the proof of the proposition. □

It then follows that the posterior distribution is $\boldsymbol{U}|\boldsymbol{Y} \sim \mathrm{B}\left(\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T\right)$ when $\boldsymbol{U}$ has a Bingham prior. Thus, the eigenvectors of $\int \boldsymbol{U}\boldsymbol{U}^T p\,(\boldsymbol{U}|\boldsymbol{Y})\, d\boldsymbol{U}$ coincide with those of $\kappa\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T + \frac{1}{2\sigma_n^2}\boldsymbol{Y}\boldsymbol{Y}^T$, with the same ordering of their eigenvalues. Consequently, the MMSD estimator is given by (13), which concludes the proof.

B Sampling from the Bingham-von Mises Fisher distribution

In this appendix, we show how to sample a unitary random matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ from a (matrix) Bingham von Mises Fisher (BMF) distribution, $\boldsymbol{X} \sim \mathrm{BMF}\,(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. As will be explained shortly, this amounts to sampling successively each column of $\boldsymbol{X}$, and entails generating a random unit norm vector drawn from a (vector) BMF distribution. We briefly review how to sample the columns of $\boldsymbol{X}$ and then explain how to sample from a vector BMF distribution.

B.1 The matrix BMF distribution

The density of $\boldsymbol{X} \sim \mathrm{BMF}\,(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is given by

$$\begin{aligned}p\,(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &\propto \mathrm{etr}\left\{\boldsymbol{C}^T\boldsymbol{X} + \boldsymbol{B}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\right\} \\ &\propto \prod_{k=1}^{p} \exp\left\{\boldsymbol{c}_k^T\boldsymbol{x}_k + \boldsymbol{B}(k,k)\boldsymbol{x}_k^T\boldsymbol{A}\boldsymbol{x}_k\right\} \qquad (30)\end{aligned}$$

where $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p\end{bmatrix}$ and $\boldsymbol{C} = \begin{bmatrix}\boldsymbol{c}_1 & \boldsymbol{c}_2 & \cdots & \boldsymbol{c}_p\end{bmatrix}$. In [34] a Gibbs-sampling strategy was presented in order to sample from this distribution, in the case where $\boldsymbol{A}$ is full-rank. We consider here a situation where $\boldsymbol{A}$ is rank-deficient and therefore we need to bring appropriate modifications to the scheme of [34] in order to handle the rank deficiency of $\boldsymbol{A}$. As evidenced from (30)

the distribution of $\boldsymbol{A}$ is a product of vector BMF distributions, except that the columns of $\boldsymbol{X}$ are not statistically independent since they are orthogonal with probability one. Let us rewrite $\boldsymbol{X}$ as $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{k-1} & \boldsymbol{Q}_\perp \boldsymbol{z} & \boldsymbol{x}_{k+1} & \cdots & \boldsymbol{x}_p\end{bmatrix}$ where $\boldsymbol{z} \in \mathcal{S}_{N-p+1} = \left\{\boldsymbol{x} \in \mathbb{R}^{N-p+1\times 1}; \boldsymbol{x}^T\boldsymbol{x} = 1\right\}$ and $\boldsymbol{Q}_\perp$ is an $N \times N-p+1$ orthonormal basis for $\mathcal{R}\left(\boldsymbol{X}_{-k}\right)^\perp$ where $\boldsymbol{X}_{-k}$ stands for the matrix $\boldsymbol{X}$ with its k-th column removed. As shown in [34] the conditional density of $\boldsymbol{z}$ given $\boldsymbol{X}_{-k}$ is

$$\begin{aligned} p\left(\boldsymbol{z}|\boldsymbol{X}_{-k}\right) &\propto \exp\left\{\boldsymbol{c}_k^T\boldsymbol{Q}_\perp\boldsymbol{z} + \boldsymbol{B}(k,k)\boldsymbol{z}^T\boldsymbol{Q}_\perp^T\boldsymbol{A}\boldsymbol{Q}_\perp\boldsymbol{z}\right\} \\ &\propto \exp\left\{\tilde{\boldsymbol{c}}_k^T\boldsymbol{z} + \boldsymbol{z}^T\tilde{\boldsymbol{A}}\boldsymbol{z}\right\} \end{aligned} \tag{31}$$

where $\tilde{\boldsymbol{c}}_k = \boldsymbol{Q}_\perp^T\boldsymbol{c}_k$ and $\tilde{\boldsymbol{A}} = \boldsymbol{B}(k,k)\boldsymbol{Q}_\perp^T\boldsymbol{A}\boldsymbol{Q}_\perp$. Therefore, $\boldsymbol{z}|\boldsymbol{X}_{-k}$ follows a vector BMF distribution $\boldsymbol{z}|\boldsymbol{X}_{-k} \sim \text{vBMF}\left(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{c}}_k\right)$. A Markov chain that converges to BMF $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ can thus be constructed as follows:

Input: initial value $\boldsymbol{X}^{(0)}$

1: **for** $k = 1, \cdots, p$ (random order) **do**
2: compute a basis $\boldsymbol{Q}_\perp$ for the null space of $\boldsymbol{X}_{-k}$ and set $\boldsymbol{z} = \boldsymbol{Q}_\perp^T\boldsymbol{x}_k$.
3: compute $\tilde{\boldsymbol{c}}_k = \boldsymbol{Q}_\perp^T\boldsymbol{c}_k$ and $\tilde{\boldsymbol{A}} = \boldsymbol{B}(k,k)\boldsymbol{Q}_\perp^T\boldsymbol{A}\boldsymbol{Q}_\perp$.
4: sample $\boldsymbol{z}$ from a vBMF $\left(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{c}}_k\right)$ distribution (*see next section*).
5: set $\boldsymbol{x}_k = \boldsymbol{Q}_\perp\boldsymbol{z}$.
6: **end for**

B.2 The vector BMF distribution

The core part of the above algorithm, see line 4, is to draw a unit-norm random vector $\boldsymbol{x}$ distributed according to a vector Bingham-von Mises Fisher distribution. The latter distribution on the M- dimensional sphere has a density with respect to the uniform distribution given by

$$p\left(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{A}\right) \propto \exp\left\{\boldsymbol{c}^T\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right\}, \boldsymbol{x} \in \mathcal{S}_M. \tag{32}$$

In [34] a Gibbs-sampling strategy was presented in order to sample from this distribution. While $\boldsymbol{A}$ was assumed to be full-rank in [34], we consider here a situation where $\boldsymbol{A}$ is rank-deficient, i.e. its eigenvalue decomposition can be written as $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T$ where $\boldsymbol{E}$ stands for the orthonormal matrix of the eigenvectors and $\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_r, 0, \cdots, 0\right)$ is the diagonal matrix of its eigenvalues. Our derivation follows along the same lines as in [34] with the appropriate modifications due to the rank deficiency of $\boldsymbol{A}$. Let $\boldsymbol{y} = \boldsymbol{E}^T\boldsymbol{x} \in \mathcal{S}_M$ and $\boldsymbol{d} = \boldsymbol{E}^T\boldsymbol{c}$. Since $y_M^2 = 1 - \sum_{k=1}^{M-1} y_k^2$, the uniform density in terms of the unconstrained coordinates $\{y_1, y_2, \cdots, y_{M-1}\}$ is proportional to $|y_M|^{-1}$ and the density of $\{y_1, y_2, \cdots, y_{M-1}\}$ is given by [34]

$$\begin{aligned} p\left(\boldsymbol{y}|\boldsymbol{d}, \boldsymbol{E}\right) &\propto \exp\left\{\boldsymbol{d}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y}\right\}|y_M|^{-1}, \quad y_M^2 = 1 - \sum_{k=1}^{M-1} y_k^2 \\ &\propto \exp\left\{\sum_{k=1}^{M} d_k y_k + \sum_{k=1}^{r} \lambda_k y_k^2\right\}|y_M|^{-1}. \end{aligned} \tag{33}$$

In order to sample from this distribution, a Gibbs sampling strategy is advocated. Towards this end, we need to derive the conditional distributions of y_k, given $\boldsymbol{y}_{-k}$ where $\boldsymbol{y}_{-k}$ stands for the vector $\boldsymbol{y}$ with its k-th component removed. Similarly to [34], let us make the change of variables $\theta_k = y_k^2$ and let $\boldsymbol{q} = \left[\frac{y_1^2}{1-y_k^2} \quad \frac{y_2^2}{1-y_k^2} \quad \cdots \quad \frac{y_M^2}{1-y_k^2}\right]^T$, so that $\left\{y_1^2, y_2^2, \cdots, y_M^2\right\} = \left\{\theta_k, (1-\theta_k)\boldsymbol{q}_{-k}\right\}$. Since this change of variables is not bijective, i.e. $y_k \pm \theta_k^{1/2}$, we need to introduce the sign s_k of y_k, and we let $\boldsymbol{s} = \begin{bmatrix}s_1 & s_2 & \cdots & s_M\end{bmatrix}^T$. Note that $y_M^2 = 1 - \sum_{k=1}^{M-1} y_k^2$,

$|y_M| = (1-\theta_k)^{1/2} q_M^{1/2}$ and $q_M = 1 - \sum_{\ell=1,\ell\neq k}^{M-1} q_\ell$. As shown in [34], the Jacobian of the transformation from $\{y_1, y_2, \cdots, y_{M-1}\}$ to $\{\theta, q_1, \cdots, q_{k-1}, q_{k+1}, \cdots, q_{M-1}\}$ is proportional to $\theta_k^{-1/2}(1-\theta_k)^{(M-2)/2}\prod_{\ell=1,\ell\neq k}^{M-1} q_\ell^{-1/2}$, and therefore the joint distribution of $\theta_k, s_k, \boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}$ can be written as

$$\begin{aligned} p\left(\theta_k, s_k, \boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) \propto \theta_k^{-1/2}(1-\theta_k)^{(M-3)/2}\left(\prod_{\ell\neq k} q_\ell^{-1/2}\right) \\ \times \exp\left\{ s_k\theta_k^{1/2} d_k + (1-\theta_k)^{1/2}\sum_{\ell\neq k} d_\ell s_\ell q_\ell^{1/2}\right\} \\ \times \begin{cases} \exp\left\{\theta_k\lambda_k + (1-\theta_k)\sum_{\ell=1,\ell\neq k}^{r} q_\ell\lambda_\ell\right\} & 1\leq k\leq r \\ \exp\left\{(1-\theta_k)\sum_{\ell=1}^{r} q_\ell\lambda_\ell\right\} & r+1\leq k\leq M \end{cases}. \end{aligned} \tag{34}$$

It follows that

- for $k \in [1, r]$

$$\begin{aligned} p\left(\theta_k, s_k|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) \propto \theta_k^{-1/2}(1-\theta_k)^{(M-3)/2}\exp\left\{\theta_k\lambda_k + (1-\theta_k)\,\boldsymbol{q}_{-k}^T\boldsymbol{\lambda}_{-k}\right\} \\ \times \exp\left\{ s_k\theta_k^{1/2} d_k + (1-\theta_k)^{1/2}\left[\boldsymbol{s}_{-k}\odot\boldsymbol{q}_{-k}^{1/2}\right]^T\boldsymbol{d}_{-k}\right\}. \end{aligned} \tag{35}$$

- for $k \in [r+1, M]$

$$\begin{aligned} p\left(\theta_k, s_k|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) \propto \theta_k^{-1/2}(1-\theta_k)^{(M-3)/2}\exp\left\{(1-\theta_k)\,\boldsymbol{q}^T\boldsymbol{\lambda}\right\} \\ \times \exp\left\{ s_k\theta_k^{1/2} d_k + (1-\theta_k)^{1/2}\left[\boldsymbol{s}_{-k}\odot\boldsymbol{q}_{-k}^{1/2}\right]^T\boldsymbol{d}_{-k}\right\}. \end{aligned} \tag{36}$$

In the previous equations, $\odot$ stands for the element-wise vector or matrix product and $\boldsymbol{q}_{-k}^{1/2}$ is a short-hand notation to designate the vector $\left[q_1^{1/2} \;\; \cdots \;\; q_{k-1}^{1/2} \;\; q_{k+1}^{1/2} \;\; \cdots \;\; q_M^{1/2}\right]^T$. In order to sample from $p\left(\theta_k, s_k|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right)$, we first sample θ_k from

$$\begin{aligned} p\left(\theta_k|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) &= p\left(\theta_k, s_k=-1|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) + p\left(\theta_k, s_k=1|\boldsymbol{q}_{-k}, \boldsymbol{s}_{-k}\right) \\ &\propto \theta_k^{-1/2}(1-\theta_k)^{(M-3)/2}\exp\left\{a_k\theta_k + b_k(1-\theta_k)^{1/2}\right\} \\ &\times \left[\exp\left\{-d_k\theta_k^{1/2}\right\} + \exp\left\{-d_k\theta_k^{1/2}\right\}\right] \end{aligned} \tag{37}$$

where $b_k = \left[\boldsymbol{s}_{-k}\odot\boldsymbol{q}_{-k}^{1/2}\right]^T\boldsymbol{d}_{-k}$ and

$$a_k = \begin{cases} \lambda_k - \boldsymbol{q}_{-k}^T\boldsymbol{\lambda}_{-k} & k\in[1,r] \\ -\boldsymbol{q}^T\boldsymbol{\lambda} & k\in[r+1,M] \end{cases}. \tag{38}$$

Next, we sample $s_k \in \{-1, +1\}$ with probabilities proportional to $\left(e^{-d_k\theta_k^{1/2}}, e^{+d_k\theta_k^{1/2}}\right)$. In order to sample from the distribution in (37), an efficient rejection sampling scheme was proposed in [34], where the proposal distribution is a beta distribution with suitably chosen parameters.

C MCMC implementation of the MMSD estimator in the covariance matrix model

In this appendix, we provide the necessary details for MCMC implementation of the MMSD estimator in the covariance matrix model. We successively investigate the case of a Bingham prior and the case of a vMF prior.

C.1 Bingham prior

When the prior distribution of $\boldsymbol{U}$ is the Bingham distribution of (9), the joint posterior distribution of $\boldsymbol{U}$ and $\boldsymbol{\gamma}$ is given by (22). Since the MMSD estimator involves the posterior distribution $p(\boldsymbol{U}|\boldsymbol{Y})$ of $\boldsymbol{U}$ only, a natural way to proceed is to marginalize (22) with respect to $\boldsymbol{\gamma}$. Let $\boldsymbol{Z} = \boldsymbol{Y}^T\boldsymbol{U} = \begin{bmatrix}\boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_p\end{bmatrix}$. Then, from (22) one has

$$\begin{aligned}
p\left(\boldsymbol{U}|\boldsymbol{Y}\right) &= \int p\left(\boldsymbol{U},\boldsymbol{\gamma}|\boldsymbol{Y}\right) d\boldsymbol{\gamma} \\
&\propto \operatorname{etr}\left\{\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U} + \frac{\nu}{2}\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{U}\right\} \\
&\times \prod_{k=1}^{p}\int_{\gamma_-}^{\gamma_+} \gamma_k^{K/2}\exp\left\{-\frac{\nu}{2}\gamma_k\left\|\boldsymbol{z}_k\right\|^2\right\} d\gamma_k \\
&\propto \operatorname{etr}\left\{\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U} + \frac{\nu}{2}\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{U}\right\} \\
&\times \prod_{k=1}^{p}\left\|\boldsymbol{z}_k\right\|^{-2(1+K/2)}\left[\gamma\left(\frac{\nu}{2}\gamma_+\left\|\boldsymbol{z}_k\right\|^2, 1+\frac{K}{2}\right) - \gamma\left(\frac{\nu}{2}\gamma_-\left\|\boldsymbol{z}_k\right\|^2, 1+\frac{K}{2}\right)\right]
\end{aligned} \tag{39}$$

where $\gamma(x,a) = \int_0^x t^{a-1}e^{-t}dt$ is the incomplete Gamma function. Unfortunately, the above distribution does not belong to any known family and it is thus problematic to generate samples drawn from it. Instead, in order to sample according to (22), we propose to use a Gibbs sampler drawing samples according to $p\left(\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\gamma}\right)$ and $p\left(\gamma_k|\boldsymbol{Y},\boldsymbol{U}\right)$ for $k = 1,\cdots,p$. From (22), the conditional distribution of $\boldsymbol{U}$ is

$$p\left(\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\gamma}\right) \propto \operatorname{etr}\left\{\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U} + \frac{\nu}{2}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{U}\right\} \tag{40}$$

which is recognized as a (modified) Bingham distribution[4]

$$\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\gamma} \sim \tilde{\mathrm{B}}\left(\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T, \kappa\boldsymbol{I}, \boldsymbol{Y}\boldsymbol{Y}^T, \frac{\nu}{2}\left(\boldsymbol{I}-\boldsymbol{\Gamma}\right)\right). \tag{41}$$

Let us now turn to the conditional distribution of $\boldsymbol{\gamma}|\boldsymbol{Y},\boldsymbol{U}$. From (22) one has

$$\begin{aligned}
p\left(\boldsymbol{\gamma}|\boldsymbol{Y},\boldsymbol{U}\right) &\propto \left|\boldsymbol{\Gamma}\right|^{K/2}\operatorname{etr}\left\{-\frac{\nu}{2}\boldsymbol{Z}\boldsymbol{\Gamma}\boldsymbol{Z}^T\right\}\left(\prod_{k=1}^{p}\mathbb{I}_{[\gamma_-,\gamma_+]}(\gamma_k)\right) \\
&\propto \prod_{k=1}^{p}\left[\gamma_k^{K/2}\exp\left\{-\frac{\nu}{2}\left\|\boldsymbol{z}_k\right\|^2\gamma_k\right\}\mathbb{I}_{[\gamma_-,\gamma_+]}(\gamma_k)\right]
\end{aligned} \tag{42}$$

which is the product of independent gamma distributions with parameters $\frac{K}{2}+1$ and $\frac{\nu}{2}\left\|\boldsymbol{z}_k\right\|^2$, truncated in the interval $[\gamma_-,\gamma_+]$. We denote this distribution as $\gamma_k \sim \mathcal{G}_t\left(\frac{K}{2}+1, \frac{\nu}{2}\left\|\boldsymbol{z}_k\right\|^2, \gamma_-, \gamma_+\right)$. Random variables with such a distribution can be efficiently generated using the accept-reject scheme of [42].

[4] $\boldsymbol{X} \sim \tilde{\mathrm{B}}\left(\boldsymbol{A}_1,\boldsymbol{B}_1,\boldsymbol{A}_2,\boldsymbol{B}_2\right) \Leftrightarrow p(\boldsymbol{X}) \propto \operatorname{etr}\left\{\boldsymbol{B}_1\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{X} + \boldsymbol{B}_2\boldsymbol{X}^T\boldsymbol{A}_2\boldsymbol{X}\right\}$

The above conditional distributions can now be used in a Gibbs sampler, as described in Algorithm 1. The so-generated matrices $\boldsymbol{U}^{(n)}$ can be used similarly to (15) to obtain the MMSD estimator.

Algorithm 1 Covariance matrix model. Gibbs sampler for the Bingham prior distribution.

Input: initial values $\boldsymbol{U}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$

1: **for** $n = 1, \cdots, N_{bi} + N_r$ **do**

2: sample $\boldsymbol{U}^{(n)}$ from $\tilde{\mathrm{B}}\left(\kappa \boldsymbol{I}, \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T, \frac{\nu}{2}\left(\boldsymbol{I} - \boldsymbol{\Gamma}^{(n-1)}\right), \boldsymbol{Y}\boldsymbol{Y}^T\right)$ in (40).

3: for $k = 1, \cdots, p$, sample $\gamma_k^{(n)}$ from $\mathcal{G}_t\left(\frac{K}{2}+1, \frac{\nu}{2}\left\|\boldsymbol{Y}^T\boldsymbol{u}_k^{(n)}\right\|^2, \gamma_-, \gamma_+\right)$ in (42).

4: **end for**

Output: sequence of random variables $\boldsymbol{U}^{(n)}$ and $\boldsymbol{\gamma}^{(n)}$

C.2 von Mises Fisher prior

When $\boldsymbol{U}$ has a vMF prior distribution, the joint posterior distribution of $\boldsymbol{U}$ and $\boldsymbol{\gamma}$ is given by (23). Marginalizing the latter with respect to $\boldsymbol{\gamma}$ will yield a posterior distribution $p(\boldsymbol{U}|\boldsymbol{Y})$ similar to that of equation (39), except that the term $\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\boldsymbol{U}$ should be replaced by $\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}}$. Again this leads to an intractable posterior. Therefore, as done previously, we consider drawing samples according to $p\left(\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{\gamma}\right)$ and $p\left(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{U}\right)$. The latter distribution will still be given by (42) while $p\left(\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{\gamma}\right)$ now takes the form

$$p\left(\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{\gamma}\right) \propto \mathrm{etr}\left\{\kappa \boldsymbol{U}^T\bar{\boldsymbol{U}} + \frac{\nu}{2}\left(\boldsymbol{I} - \boldsymbol{\Gamma}\right)\boldsymbol{U}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{U}\right\} \tag{43}$$

which is recognized as a BMF distribution $\boldsymbol{U}|\boldsymbol{Y}, \boldsymbol{\gamma} \sim \mathrm{BMF}\left(\boldsymbol{Y}\boldsymbol{Y}^T, \frac{\nu}{2}\left(\boldsymbol{I} - \boldsymbol{\Gamma}\right), \kappa\bar{\boldsymbol{U}}\right)$. Therefore only line 2 of the Gibbs sampler in Table 1 needs to be modified, which yields the Gibbs sampler of Algorithm 2 in the case of a vMF prior.

Algorithm 2 Covariance matrix model. Gibbs sampler for the vMF prior distribution.

Input: initial values $\boldsymbol{U}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$

1: **for** $n = 1, \cdots, N_{bi} + N_r$ **do**

2: sample $\boldsymbol{U}^{(n)}$ from $\mathrm{BMF}\left(\boldsymbol{Y}\boldsymbol{Y}^T, \frac{\nu}{2}\left(\boldsymbol{I} - \boldsymbol{\Gamma}^{(n-1)}\right), \kappa\bar{\boldsymbol{U}}\right)$ in (43).

3: for $k = 1, \cdots, p$, sample $\gamma_k^{(n)}$ from $\mathcal{G}_t\left(\frac{K}{2}+1, \frac{\nu}{2}\left\|\boldsymbol{Y}^T\boldsymbol{u}_k^{(n)}\right\|^2, \gamma_-, \gamma_+\right)$ in (42).

4: **end for**

Output: sequence of random variables $\boldsymbol{U}^{(n)}$ and $\boldsymbol{\gamma}^{(n)}$

Acknowledgment

The authors would like to thank Prof. Kit Bingham from the University of Minnesota for insightful comments on the Bingham distribution and for pointing reference [43]. They are also indebted to Prof. Gérard Letac, University of Toulouse, for fruitful discussions leading to the proof of Proposition 2 given in Appendix A.

References

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

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions Antennas Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[3] R. Roy and T. Kailath, "ESPRIT- estimation of signal parameters via rotational invariance techniques," *IEEE Transactions Acoustics Speech Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[4] 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.

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

[6] B. Ottersten, M. Viberg, P. Stoica, and A. Nehorai, "Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing," in *Radar Array Processing*, S. Haykin, J. Litva, and T. Shepherd, Eds. Berlin: Springer Verlag, 1993, ch. 4, pp. 99–151.

[7] 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.

[8] ——, "MUSIC, maximum likelihood and Cramér-Rao bound: Further results and comparisons," *IEEE Transactions Acoustics Speech Signal Processing*, vol. 38, no. 12, pp. 2140–2150, December 1990.

[9] 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.

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

[11] B. Johnson, Y. Abramovich, and X. Mestre, "The role of subspace swap in MUSIC performance breakdown," in *Proceedings ICASSP*, March 2008, pp. 2473 –2476.

[12] R. R. Nadakuditi and F. Benaych-Georges, "The breakdown point of signal subspace estimation," in *Proceedings SAM*, Israel, 4-7 October 2010, pp. 177–180.

[13] 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.

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

[15] 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.

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

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

[18] U. Grenander, M. I. Miller, and A. Srivastava, "Hilbert-Schmidt lower bounds for estimators on matrix Lie groups for ATR," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 20, no. 8, pp. 790–802, August 1998.

[19] Q. Rentmeesters, P. Absil, P. Van Dooren, K. Gallivan, and A. Srivastava, "An efficient particle filtering technique on the Grassmann manifold," in *Proceedings ICASSP*, Dallas, TX, March 14-19 2010, pp. 3838–3841.

[20] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton, NJ: Princeton University Press, 2008.

[21] P.-A. Absil, "Optimization on manifolds: Methods and applications," Université Catholique Louvain, Tech. Rep. UCL-INMA-2009.043, 2009.

[22] T. E. Abrudan, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Transactions Signal Processing*, vol. 56, no. 3, pp. 1134–1147, March 2008.

[23] ——, "Conjugate gradient algorithm for optimization under unitary matrix constraint," *Signal Processing*, vol. 89, pp. 1704–1714, 2009.

[24] S. Fiori and T. Tanaka, "An algorithm to compute averages on matrix lie groups," *IEEE Transactions Signal Processing*, vol. 57, no. 12, pp. 4734–4743, December 2009.

[25] J. J. Boutros, F. Kharrat-Kammoun, and H. Randriambololona, "A classification of multiple antenna channels," in *Proceedings International Zurich Seminar on Communications*, February 22-24 2006, pp. 14–17.

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

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

[28] K. Mammasis, R. W. Stewart, and J. S. Thompson, "Spatial fading correlation model using mixtures of von Mises Fisher distributions," *IEEE Transactions Wireless Communications*, vol. 8, no. 4, pp. 2046–2055, April 2009.

[29] K. Mammasis and R. W. Stewart, "The Fisher-Bingham spatial correlation model for multielement antenna systems," *IEEE Transactions Vehicular Technology*, vol. 58, no. 5, pp. 2130–2136, June 2009.

[30] O. Hamsici and A. Martinez, "Rotation invariant kernels and their application to shape analysis," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 31, no. 11, pp. 1985–1999, November 2009.

[31] J. Ward, "Space-time adaptive processing for airborne radar," Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, Tech. Rep. 1015, December 1994.

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

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

[34] 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.

[35] A. Sarlette and R. Sepulchre, "Consensus optimization on manifolds," *SIAM Journal Control Optimization*, vol. 48, no. 1, pp. 56–76, 2009.

[36] E. Begelfor and M. Werman, "Affine invariance revisited," in *Proceedings IEEE CVPR06*, 2006, pp. 2087–2094.

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

[38] A. Halimi, Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Transactions Geoscience Remote Sensing*, 2011, to appear.

[39] W. Fan, B. Hu, J. Miller, and M. Li, "Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data," *International Journal Remote Sensing*, vol. 30, no. 11, pp. 2951–2962, June 2009.

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

[41] G. Letac, "Familles exponentielles invariantes sur les matrices symétriques," December 2010, private communication.

[42] Y. Chung, "Simulation of truncated gamma variables," *Korean J. Comput. Appl. Math.*, vol. 5, no. 3, pp. 601–610, 1998.

[43] P. E. Jupp and K. V. Mardia, "Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions," *The Annals of Statistics*, vol. 7, no. 3, pp. 599–606, May 1979.

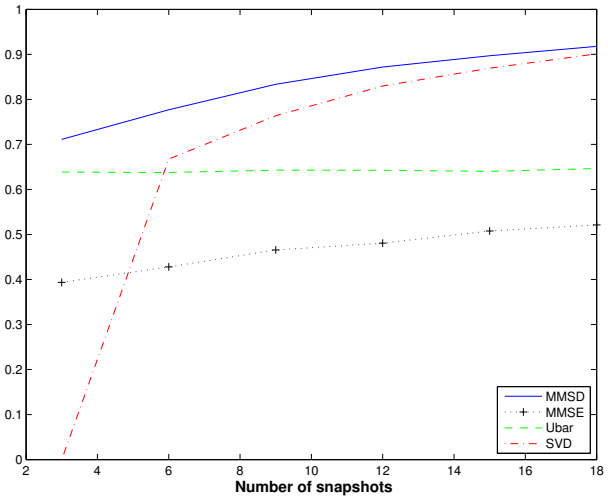

Figure 4: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus K. $N = 20$, $p = 5$, $\kappa = 20$ and $SNR = 5$dB. Linear model, Bingham prior.

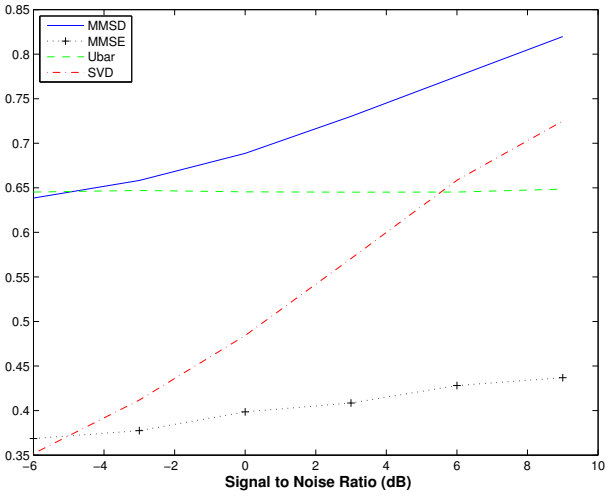

Figure 5: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus SNR. $N = 20$, $p = 5$, $\kappa = 20$ and $K = 5$. Linear model, Bingham prior.

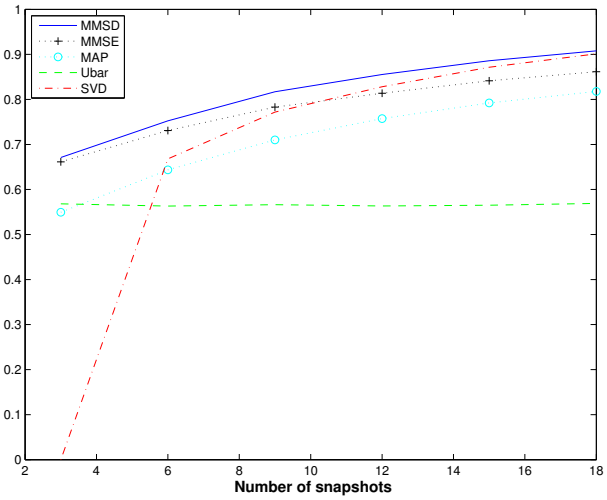

Figure 6: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus K. $N = 20$, $p = 5$, $\kappa = 20$ and $SNR = 5$dB. Linear model, vMF prior.

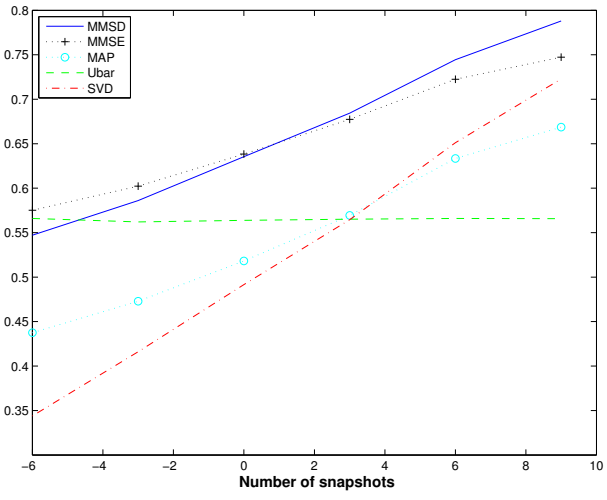

Figure 7: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus SNR. $N = 20$, $p = 5$, $\kappa = 20$ and $K = 5$. Linear model, vMF prior.

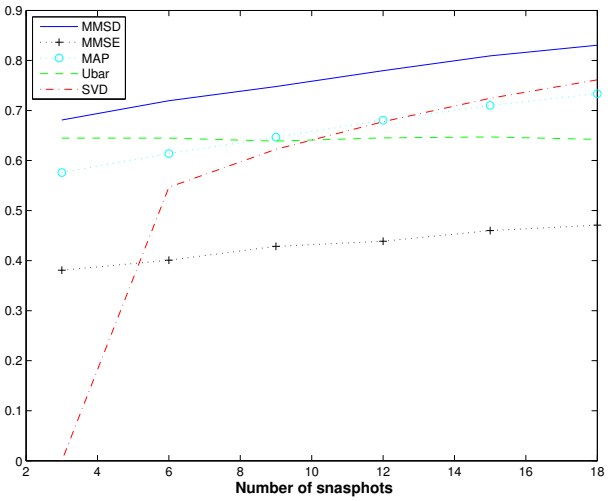

Figure 8: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus K. $N = 20$, $p = 5$, $\kappa = 20$, $SNR_{-} = 5$dB and $SNR_{+} = 10$dB. Covariance matrix model, Bingham prior.

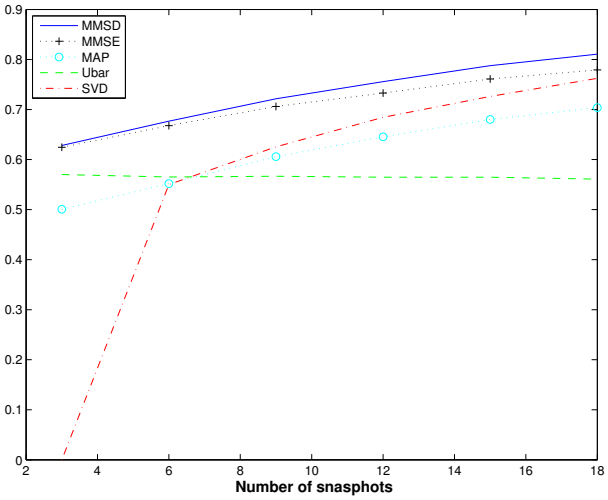

Figure 9: Fraction of energy of $\hat{\boldsymbol{U}}$ in $\mathcal{R}(\boldsymbol{U})$ versus K. $N = 20$, $p = 5$, $\kappa = 20$, $SNR_{-} = 5$dB and $SNR_{+} = 10$dB. Covariance matrix model, vMF prior.

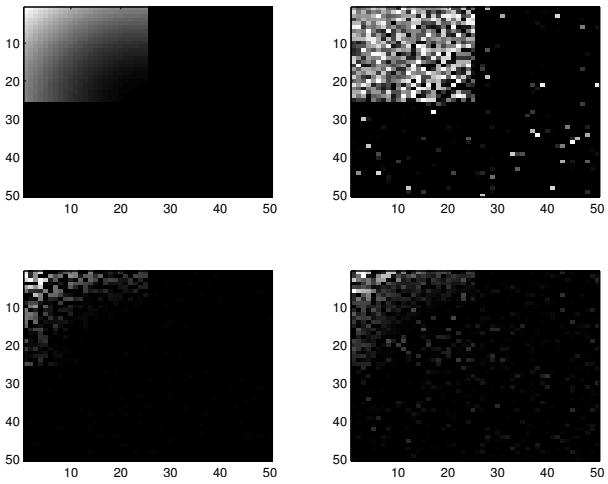

Figure 10: Top, left: non-linearity coefficients $\gamma_{1,2}$. Top, right: distance between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_n$ estimated with $\eta = 0$. Bottom: distance between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_\ell$ estimated with $\eta = 0.5$ (left) and $\eta = 50$ (right).

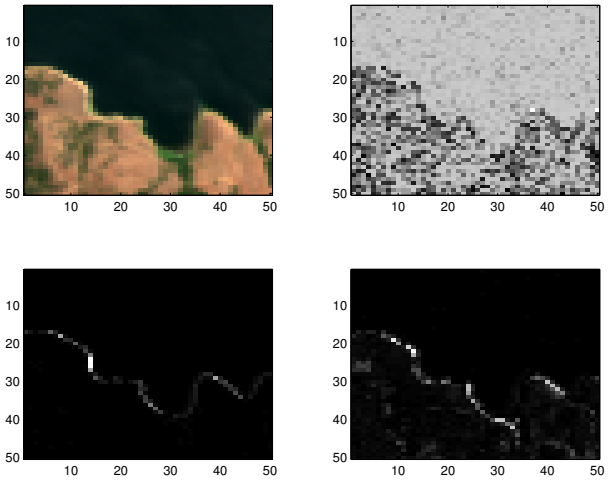

Figure 11: Top, left: The Moffett Field scene as composite true colors. Top, right: distance between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_n$ estimated with $\eta = 0$. Bottom: distance between $\bar{\boldsymbol{U}}$ and $\hat{\boldsymbol{U}}_\ell$ estimated with $\eta = 0.5$ (left) and $\eta = 50$ (right).