Joint Bayesian Estimation of Close Subspaces from Noisy Measurements

Olivier Besson, *Senior Member, IEEE*, Nicolas Dobigeon, *Member, IEEE*, and Jean-Yves Tourneret, *Senior Member, IEEE*

***Abstract*—In this letter, we consider two sets of observations defined as subspace signals embedded in noise and we wish to analyze the distance between these two subspaces. The latter entails evaluating the angles between the subspaces, an issue reminiscent of the well-known Procrustes problem. A Bayesian approach is investigated where the subspaces of interest are considered as random with a joint prior distribution (namely a Bingham distribution), which allows the closeness of the two subspaces to be parameterized. Within this framework, the minimum mean-square distance estimator of both subspaces is formulated and implemented via a Gibbs sampler. A simpler scheme based on alternative maximum a posteriori estimation is also presented. The new schemes are shown to provide more accurate estimates of the angles between the subspaces, compared to singular value decomposition based independent estimation of the two subspaces.**

***Index Terms*—Bingham distribution, Procrustes problem, subspace estimation.**

I. Problem Statement

MODELING signals of interest as belonging to a linear subspace is arguably one of the most encountered approach in engineering applications [1]–[3]. Estimation of such signals in additive white noise is usually conducted via the singular value decomposition which has proven to be very successful in numerous problems, including spectral analysis or direction finding. In this letter, we consider a situation where two independent noisy observations of a subspace signal are available but, due to miscalibration or a change in the observed process, the subspace of interest is slightly different from one observation to the other. More precisely, assume that we observe two $M \times T$ matrices X_1 and X_2 given by

$$X_k = H_k S_k + N_k; \quad k = 1, 2 \tag{1}$$

where the orthogonal $M \times R$ matrices H_k ($H_k^T H_k = I_R$) span the subspace where the signals of interest lie, S_k stands for the matrix of coordinates of the noise-free data within the range space $\mathcal{R}(H_k)$ of H_k, and N_k denotes an additive white Gaussian noise. Herein, we are interested in recovering the subspaces H_1, H_2 but, maybe *more importantly*, to have an indication of the "difference" between these two subspaces. The natural distance between H_1 and H_2 is given by $\left[\sum_{r=1}^R \theta_r^2\right]^{1/2}$ where θ_r are the principal angles between H_1 and H_2, which can be obtained from the singular value decomposition (SVD) $H_1^T H_2 = Y \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_R) Z^T$. This problem is somehow reminiscent of the orthogonal matrix Procrustes problem [4, p. 601] where one seeks an orthogonal matrix that brings H_1 close to H_2 by solving $\min_{Q^T Q = I} \|H_2 - H_1 Q\|_F$. The solution is well known to be $Q = Y Z^T$. The problem here is slightly different as we only have access to X_1, X_2 and not to the subspaces themselves. Moreover, we would like to exploit the fact that H_1 and H_2 are close subspaces. In order to embed this knowledge, a Bayesian framework is formulated where H_1 and H_2 are treated as random matrices with a joint distribution, as detailed now.

Let us state our assumptions and our approach to estimating H_1, H_2 and subsequently the principal angles $\theta_r, r = 1, \cdots, R$. Assuming that the columns of N_1 and N_2 are independent and identically Gaussian distributed $N_k \sim \mathcal{N}(0, \sigma^2 I)$ with σ^2 known, the likelihood function of X_k is given by

$$p(X_k | H_k, S_k) \propto \operatorname{etr}\left\{-\frac{1}{2\sigma^2}(X_k - H_k S_k)^T (X_k - H_k S_k)\right\} \tag{2}$$

where $\propto$ means proportional to and $\operatorname{etr}\{.\}$ stands for the exponential of the trace of the matrix between braces. As for S_k, we assume that no knowledge about it is available so that its prior distribution is given by $\pi(S_k) \propto 1$. Note that this is an improper prior but, as will be shown shortly, marginalizing with respect to S_k results in a proper distribution. Indeed,

$$\begin{aligned} p(X_k | H_k) &= \int p(X_k | H_k, S_k) \pi(S_k) dS_k \\ &\propto \operatorname{etr}\left\{-\frac{1}{2\sigma^2}(X_k^T X_k - X_k^T H_k H_k^T X_k)\right\}. \end{aligned} \tag{3}$$

Let us now turn to our assumption regarding H_1 and H_2. We assume that H_1 is uniformly distributed on the Stiefel manifold [5] and that H_2, conditioned on H_1, follows a Bingham distribution [5], [6] with parameter matrix $\kappa H_1 H_1^T$, i.e.,

$$\pi(H_2 | H_1) = C(\kappa H_1 H_1^T) \operatorname{etr}\{\kappa H_2^T H_1 H_1^T H_2\} \tag{4}$$

where $C(A) = {}_1F_1\left(\frac{R}{2}, \frac{N}{2}; A\right)$ and ${}_1F_1(., .; .)$ is an hypergeometric function of matrix argument [5]. It is known that ${}_1F_1(p, q; A)$ depends only on the non-zero eigenvalues of A: hence $C(\kappa H_1 H_1^T)$ in (4) depends on κ only. The latter rules the prior distribution of the angles between $\mathcal{R}(H_1)$ and $\mathcal{R}(H_2)$: the larger κ the closer $\mathcal{R}(H_1)$ and $\mathcal{R}(H_2)$ [7].

Manuscript received October 01, 2013; revised December 04, 2013; accepted December 17, 2013. Date of current version December 26, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ba-Tuong Vo.

O. Besson is with the ISAE, Department Electronics Optronics Signal, University of Toulouse, 31055 Toulouse, France (e-mail: olivier.besson@isae.fr).

N. Dobigeon and J.-Y. Tourneret are with the IRIT/ENSEEIHT, University of Toulouse, 31071 Toulouse, France (e-mail: nicolas.dobigeon@enseeiht.fr;jean-yves.tourneret@enseeiht.fr).

II. Subspace Estimation

Our objective is, given the likelihood function in (3) and the prior in (4), to estimate H_1, H_2 and then deduce the principal angles between them. Towards this end, let us first write the joint posterior distribution of H_1 and H_2 as

$$\begin{aligned} p(H_1, H_2|X_1, X_2) &\propto p(X_1, X_2|H_1, H_2)\pi(H_2|H_1)\pi(H_1) \\ &\propto \operatorname{etr}\left\{\frac{1}{2\sigma^2}X_1^T H_1 H_1^T X_1 + \frac{1}{2\sigma^2}X_2^T H_2 H_2^T X_2\right\} \\ &\times \operatorname{etr}\{\kappa H_2^T H_1 H_1^T H_2\}. \end{aligned} \tag{5}$$

In the sequel we let $\bar{k} = \{1, 2\}\backslash k$. The posterior density of H_k only is thus

$$\begin{aligned} p(H_k|X_1, X_2) &= \int p(H_k, H_{\bar{k}}|X_1, X_2)dH_{\bar{k}} \\ &\propto \operatorname{etr}\left\{\frac{1}{2\sigma^2}X_k^T H_k H_k^T X_k\right\} \\ &\times \int \operatorname{etr}\left\{H_{\bar{k}}^T\left[\frac{1}{2\sigma^2}X_{\bar{k}}X_{\bar{k}}^T + \kappa H_k H_k^T\right]H_{\bar{k}}\right\}dH_{\bar{k}} \\ &\propto C\left(\frac{1}{2\sigma^2}X_{\bar{k}}X_{\bar{k}}^T + \kappa H_k H_k^T\right)\operatorname{etr}\left\{\frac{1}{2\sigma^2}X_k^T H_k H_k^T X_k\right\}. \end{aligned} \tag{6}$$

The minimum mean-square distance (MMSD) estimator of H_k is defined as [7]

$$\begin{aligned} \hat{H}_{k-\text{MMSD}} &= \arg\min_{\hat{H}_k} \mathcal{E}\left\{\|\hat{H}_k\hat{H}_k^T - H_k H_k^T\|^2\right\} \\ &= \mathcal{P}_R\left\{\int H_k H_k^T p(H_k|X_1, X_2)dH_k\right\} \end{aligned} \tag{7}$$

where $\mathcal{E}\{.\}$ is the statistical mean and $\mathcal{P}_R\{\cdot\}$ stands for the R principal eigenvectors of the matrix between braces. From inspection of $p(H_k|X_1, X_2)$, the above integral in (7) does not seem to be tractable. Therefore, we turn to Markov chain Monte-Carlo (MCMC) simulation methods to approximate it [8]. However, the distribution in (6) is not obvious to sample. On the contrary, the conditional distribution of $H_k|H_{\bar{k}}, X_1, X_2$ belongs to a known family. Indeed, from (5) one has

$$p(H_k|H_{\bar{k}}, X_1, X_2) \propto \operatorname{etr}\left\{H_k^T\left[\frac{1}{2\sigma^2}X_k X_k^T + \kappa H_{\bar{k}}H_{\bar{k}}^T\right]H_k\right\} \tag{8}$$

which is recognized as a Bingham distribution, i.e.,

$$H_k|H_{\bar{k}}, X_1, X_2 \sim \mathrm{B}\left(\frac{1}{2\sigma^2}X_k X_k^T + \kappa H_{\bar{k}}H_{\bar{k}}^T\right). \tag{9}$$

This leads us to consider a Gibbs sampling scheme which uses (9) to draw samples asymptotically distributed according to $p(H_k|X_1, X_2)$. An efficient scheme to draw random matrices from a Bingham distribution can be found in [9]. Our Gibbs sampling scheme is summarized in Table I

Once a set of N_r matrices $H_1(n)$ and $H_2(n)$ has been generated, the MMSD estimator of H_k can be approximated as

$$\hat{H}_{k-\text{MMSD}} = \mathcal{P}_R\left\{N_\text{r}^{-1}\sum_{n=N_\text{bi}+1}^{N_\text{bi}+N_\text{r}} H_k(n)H_k(n)^T\right\}. \tag{10}$$

TABLE I
GIBBS SAMPLER FOR ESTIMATION OF H_1 AND H_2

Input: initial value $H_1(0)$
1: **for** $n = 1, \cdots, N_\text{bi} + N_\text{r}$ **do**
2: sample $H_2(n)$ from $\mathrm{B}\left(\frac{1}{2\sigma^2}X_2X_2^T + \kappa H_1(n-1)H_1(n-1)^T\right)$.
3: sample $H_1(n)$ from $\mathrm{B}\left(\frac{1}{2\sigma^2}X_1X_1^T + \kappa H_2(n)H_2(n)^T\right)$.
4: **end for**
Output: sequence of random matrices $H_1(n)$ and $H_2(n)$.

TABLE II
ITERATIVE MAP ESTIMATION OF H_1 AND H_2

Input: initial value $H_1(0)$
1: **for** $n = 1, \cdots, N_\text{it}$ **do**
2: evaluate $H_2(n) = \mathcal{P}_R\left\{\frac{1}{2\sigma^2}X_2X_2^T + \kappa H_1(n-1)H_1(n-1)^T\right\}$.
3: evaluate $H_1(n) = \mathcal{P}_R\left\{\frac{1}{2\sigma^2}X_1X_1^T + \kappa H_2(n)H_2(n)^T\right\}$.
4: **end for**
Output: $\hat{H}_{k-\text{MAP}} = H_k(N_\text{it})$.

We should point out that the scheme of Table I is computationally intensive, due to the need to generate matrices from a Bingham distribution, and that it may be prohibitive in large-scale problems when M is large. In such cases, one might turn to simpler estimators.

An alternative and possibly more computationally efficient approach would entail considering maximum a posteriori (MAP) estimation. However, the joint MAP estimation of H_1 and H_2 from $p(H_1, H_2|X_1, X_2)$ in (5) does not appear tractable. It is in fact customary in this case to consider iterative alternate maximization of $p(H_1, H_2|X_1, X_2)$, i..e, maximize it first with respect to H_1 holding H_2 fixed, and then with respect to H_2 holding H_1 fixed. Convergence of this method to the global maximum is yet to be proven, although we did not experiment problems in our simulations. At each step, the MAP estimation of one matrix, conditioned on the other one, is simple as

$$\begin{aligned} \hat{H}_{k-\text{MAP}}|H_{\bar{k}} &= \arg\max_{H_k} p(H_k|H_{\bar{k}}, X_1, X_2) \\ &= \mathcal{P}_R\left\{\frac{1}{2\sigma^2}X_k X_k^T + \kappa H_{\bar{k}}H_{\bar{k}}^T\right\}. \end{aligned} \tag{11}$$

Note that (11) is also the MMSD estimator of H_k given $H_{\bar{k}}$ since, if $H \sim \mathrm{B}(A)$, the MMSD estimator of H is simply $\mathcal{P}_R\{A\}$ [7]. Therefore we propose the scheme of Table II which we refer to as iterative MAP (iMAP).

Remark 1. (Estimation by Regularization): We have decided in this work to embed the knowledge that $\mathcal{R}(H_1)$ is close to $\mathcal{R}(H_2)$ in a prior distribution. An alternative would be to consider regularized maximum likelihood estimation (MLE). Such an approach would amount to consider the following optimization problem:

$$\begin{aligned} \min_{H_1, H_2, S_1, S_2} &-\log p(X_1, X_2|H_1, H_2, S_1, S_2) \\ &+ \mu\|H_1H_1^T - H_2H_2^T\|_F^2. \end{aligned} \tag{12}$$

Solving for S_1, S_2 and concentrating the criterion, one ends up with minimizing

$$J(H_1, H_2) = \mathrm{Tr}\left\{\frac{1}{2\sigma^2} X_1^T H_1 H_1^T X_1\right\} + \mathrm{Tr}\left\{\frac{1}{2\sigma^2} X_2^T H_2 H_2^T X_2\right\} + \mathrm{Tr}\{2\mu H_2^T H_1 H_1^T H_2\}. \tag{13}$$

From observation of (5) this is tantamount to maximizing $p(H_1, H_2|X_1, X_2)$ with the regularization parameter 2μ playing a similar role as κ. However, there are two differences. First, in a Bayesian setting κ can be fixed by looking at the prior distribution of the angles between $\mathcal{R}(H_1)$ and $\mathcal{R}(H_2)$ and making it match our prior knowledge. Second, the Bayesian framework enables one to consider an MMSD estimator while the frequentist approach bears much resemblance with a maximum a posteriori estimator.

Remark 2. (Alternative Prior Modeling): Instead of considering a Bingham distribution as prior for $\pi(H_2|H_1)$ a von Mises-Fisher (vMF) distribution [6] defined as

$$\pi(H_2|H_1) \propto \mathrm{etr}\{cH_2^T H_1\} \tag{14}$$

might have been used. Under this hypothesis, it is straightforward to show that the conditional posterior distribution $p(H_k|H_{\bar{k}}, X_1, X_2)$ is now Bingham von Mises-Fisher (BMF). The Gibbs sampling scheme needs to be adapted to these new distributions. However, for a BMF distribution, there does not exist a closed-form expression for the MAP estimator which means that the iterative scheme of Algorithm II cannot be extended.

Remark 3. (Extension to More than 2 Subspaces): Let us consider a situation where $K > 2$ data matrices $X_k = H_k S_k + N_k$ are available, so that their joint distribution, conditioned on $H_{1\ldots K}$ can be written as

$$p(X_{1\ldots K}|H_{1\ldots K}) \propto \mathrm{etr}\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K}(X_k^T X_k - X_k^T H_k H_k^T X_k)\right\}. \tag{15}$$

Let us still assume that H_1 is uniformly distributed on the Stiefel manifold and that for $k > 2$, $H_k|H_{k-1} \sim \mathrm{B}(\kappa_k H_{k-1} H_{k-1}^T)$. Then the joint posterior distribution of $H_{1\ldots K}$ writes

$$p(H_{1\ldots K}|X_{1\ldots K}) \propto \mathrm{etr}\left\{\frac{1}{2\sigma^2}\sum_{k=1}^{K} X_k^T H_k H_k^T X_k\right\} \times \mathrm{etr}\left\{\sum_{k=2}^{K} \kappa_k H_k^T H_{k-1} H_{k-1}^T H_k\right\}. \tag{16}$$

It ensues that the conditional posterior distribution of H_k is given by

$$H_1|H_{2\ldots K}, X_{1\ldots K} \sim \mathrm{B}\left(\frac{1}{2\sigma^2} X_1 X_1^T + \kappa_2 H_2 H_2^T\right) \tag{17a}$$

$$H_k|H_{-k}, X_{1\ldots K} \sim \mathrm{B}\left(\frac{1}{\sigma^2} X_k X_k^T + \kappa_k H_{k-1} H_{k-1}^T\right). \tag{17b}$$

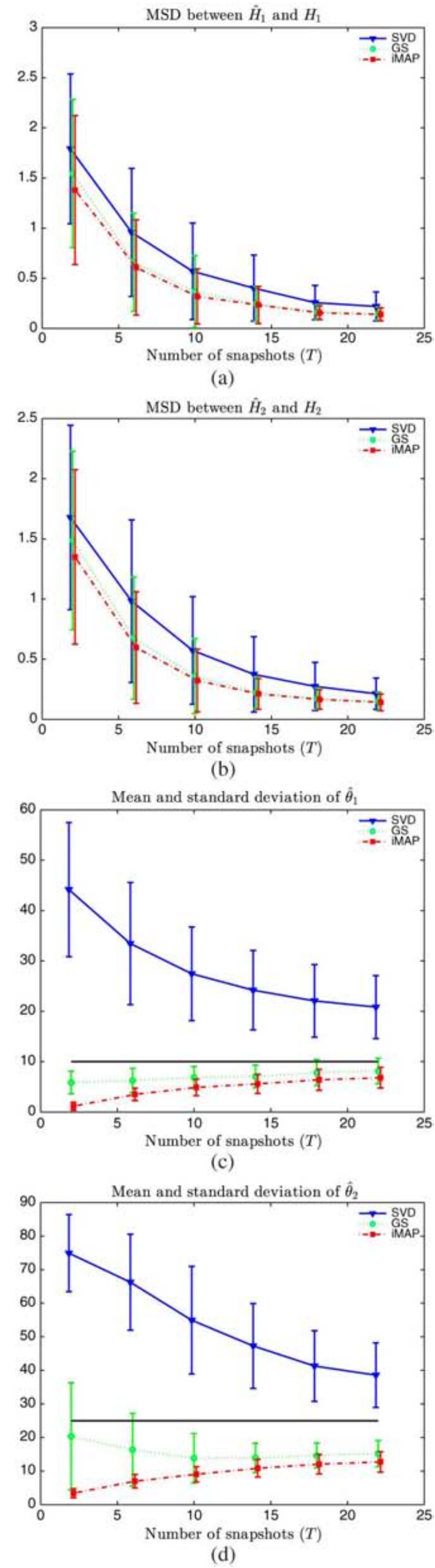

Fig. 1. Performance of the estimators versus T. $\kappa = 40$ and $\mathrm{SNR} = 0$ dB. (a) $MSD(\hat{H}_1, H_1)$, (b) $MSD(\hat{H}_2, H_2)$, (c), mean and std of $\hat{\theta}_1$, (d), mean and std of $\hat{\theta}_2$.

The Gibbs sampling scheme of Table I as well as the iterative MAP algorithm of Table II can be straightforwardly modified so as to account for this more general setting.

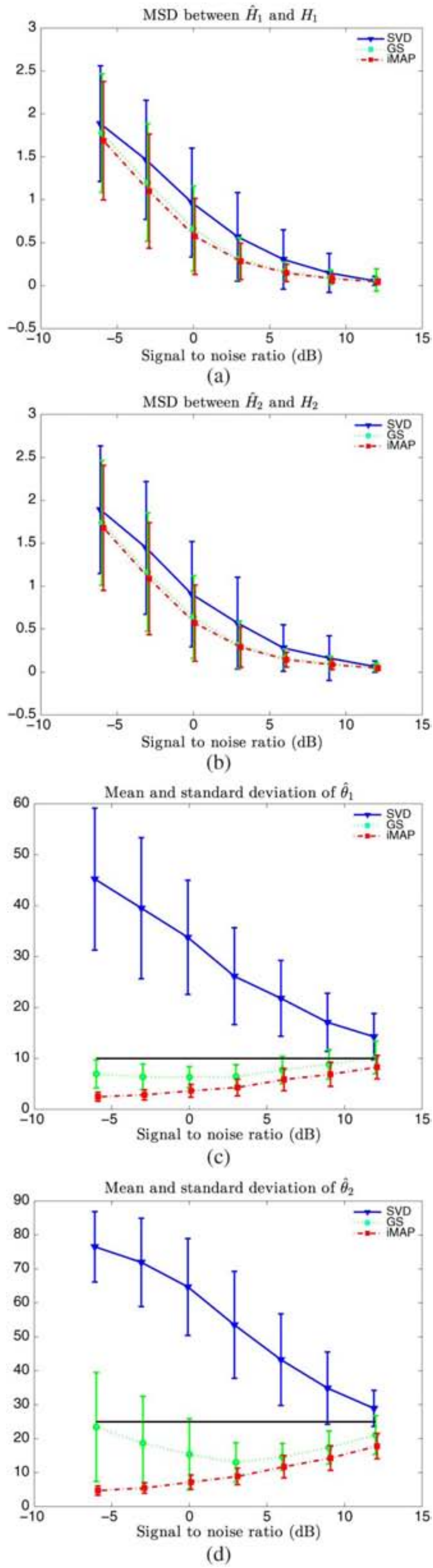

Fig. 2. Performance of the estimators versus SNR. $\kappa = 40$ and $T = 6$. (a) $MSD(\hat{H}_1, H_1)$, (b) $MSD(\hat{H}_2, H_2)$, (c), mean and std of $\hat{\theta}_1$, (d), mean and std of $\hat{\theta}_2$.

III. Numerical Illustrations

Let us now give some illustrative examples about the estimators developed above. We consider a scenario with $M = 8$ and $R = 2$. The two algorithms described above (referred to as GS and iMAP in the figures, respectively) will be compared to a conventional SVD-based approach where H_k is estimated from the R dominant left singular vectors of the data matrix X_k. For each algorithm, the angles between H_1 and H_2 will be estimated from the singular value decomposition of $\hat{H}_1^T \hat{H}_2$, where $\hat{H}_1, \hat{H}_2$ stand for one of the three estimates mentioned previously. Two criteria will be used to assess the performance of the estimators. First, the MSD between $\hat{H}_k$ and H_k will be used: this gives an idea of how accurately each subspace individually is estimated. Next, since the difference between H_1 and H_2 is of utmost importance, we will also pay attention to the mean and standard deviation of $\hat{\theta}_r$ as these angles characterize how H_2 has been moved apart from H_1.

In all simulations the entries of S_1 and S_2 were generated as i.i.d. $\mathcal{N}(0, 1)$. The subspaces H_1 and H_2 were fixed and the *true* angles between them are equal to $10°$ and $25°$ respectively. Note that the subspaces H_1 and H_2 are not generated according to the prior distributions assumed above. The signal to noise ratio (SNR) is defined as $\mathrm{SNR} = \sigma^{-2} M^{-1} R$. For the Bayesian estimators, we set $N_{\mathrm{bi}} = 10$, $N_{\mathrm{r}} = 200$ and $N_{\mathrm{it}} = 50$. In Fig. 1 we plot the performance versus T, for $\kappa = 40$, while Fig. 2 studies the performance versus SNR. The following observations can be made:

- The Bayesian estimates of the individual subspaces outperform the SVD-based estimates, especially for a small number of snapshots or a low SNR. When SNR increases however, the SVD-based estimates produce accurate estimates of each subspace.
- The SVD-based estimator does not accurately estimate the angles between H_1 and H_2, unless SNR is large. In contrast, the Bayesian estimators provide a rather accurate estimation of θ_r.
- The Gibbs sampler is seen to perform better that the iMAP estimator, at the price of a larger computational cost however.

References

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

[2] L. Scharf, "The SVD and reduced rank signal processing," *Signal Process.*, vol. 25, no. 2, pp. 113–133, Nov. 1991.

[3] A. Van der Veen, E. Deprettere, and A. Swindlehurst, "Subspace-based signal analysis using Singular Value Decomposition," *Proc. IEEE*, vol. 81, no. 9, pp. 1277–1308, Sep. 1993.

[4] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: John Hopkins Univ. Press, 1996.

[5] Y. Chikuse, *Statistics on Special Manifolds*. New York, NY, USA: Springer Verlag, 2003.

[6] K. V. Mardia and P. E. Jupp, *Directional Statistics*. New York, NY, USA: Wiley , 1999.

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

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

[9] P. D. Hoff, "Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data," *J. Comput. Graph. Statist.*, vol. 18, no. 2, pp. 438–456, Jun. 2009.