On the Latent Information Geometry of the Grassmann Manifold

Lorenzo Cazzella[1] **Søren Hauberg**[2] **Georgios Arvanitidis**[2] **Matteo Matteucci**[1]
[1] Politecnico di Milano, Department of Electronics Information and Bioengineering, Milan, Italy
[2] Technical University of Denmark, Section for Cognitive Systems, Lyngby, Denmark

Abstract

Modeling linear subspaces and relations among them naturally arises in several applications in signal processing, computer vision, and system identification. In this paper, we investigate the latent information geometry of deep generative models that output linear subspaces. Such subspaces are members of the Grassmann manifold, which we model with a matrix Bingham distribution as a likelihood. We derive the Fisher-Rao metric on the statistical manifold of the matrix Bingham parameters, and propose pulling this back to the latent space to achieve uncertainty-aware and identifiable latent representations. We provide numerical results assessing the meaningfulness of the achieved latent subspace representations on a relevant vehicular wireless communications scenario.

1 MOTIVATION

This work is motivated by a key challenge in wireless communication systems, such as cellular mobile networks. Such systems model how electromagnetic (EM) waves propagate, e.g., in urban environments, to increase both reliability and performance. In this regime, data is both limited and noisy due to the high mobility, e.g., a cell phone operated from a moving vehicle. Knowledge of how the EM waves propagate can play a significant role in helping the receiver denoise this limited data. A simple, yet powerful, approach is to project the high-dimensional data onto a lower-dimensional subspace, which is *location-related* (Brighente et al., 2020; Mizmizi et al., 2021).

We, thus, need a model of continuously changing subspaces. For this, we target a deep generative model that emits subspaces. We can then devise a control mechanism for the communication system that continuously changes a low dimensional latent variable of the generative model. Unfortunately, such latent variables are not identifiable (Syrota et al., 2025), which notably complicates the control system.

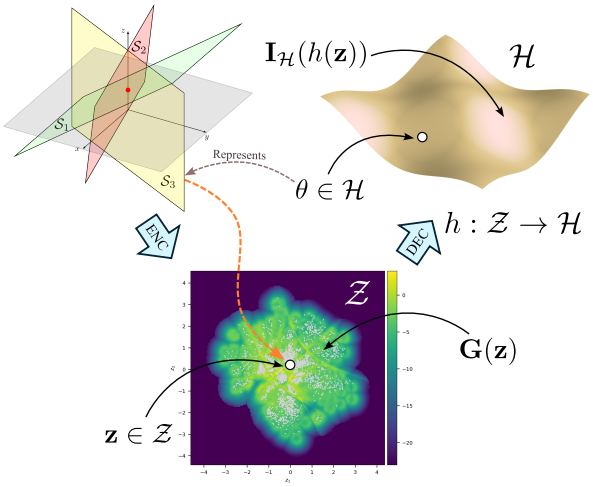


Figure 1: We explore the latent representation of a set of linear subspaces lying on a Grassmann manifold, leveraging *information geometry* to provide the latent space with an uncertainty-aware pull-back metric, whose volume is represented by the color map in the latent space $\mathcal{Z}$. Here, $\mathbf{z} \in \mathcal{Z}$ is the latent variable, $\mathbf{G}$ is the learned metric, $\mathcal{H}$ is the parameter space of the matrix Bingham distribution, and $\mathbf{I}_{\mathcal{H}}$ is the Fisher-Rao metric associated with this space.

These considerations lead to two challenges. First, we need a high-quality generative model that emits subspaces. Conventional VAE architectures aim to model high-dimensional data in a Euclidean ambient space, while in our case the output space is structured as the space of subspaces of fixed dimension, that is, the Grassmann manifold. Second, we need a representative low-dimensional embedding space on which to operate more efficiently. Arvanitidis et al. (2022) have shown that probabilistic latent-variable models

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

can be suitably endowed with a latent Riemannian structure derived by pulling back the information geometry of a meaningful distribution parameters space. Such models provide uncertainty-aware latent variables and effectively solve identifiability under strict guarantees (Syrota et al., 2025). We can then leverage the latent Riemannian structure to compute meaningful latent geodesics, which we can use, e.g., for latent data interpolation or clustering. This requires us to understand the information geometry of a representative probability distribution over the Grassmann manifold.

In this paper, we investigate the latent representation of linear subspaces lying on the Grassmann manifold, leveraging *information geometry* to achieve uncertainty-aware and identifiable latent representations. First, we construct a probabilistic deep latent variable model over subspaces by selecting the matrix Bingham distribution as a likelihood owing to its invariance properties (Sec. 3). Then, we explore the latent information geometry of the Grassmann manifold (Sec. 4), deriving novel computationally tractable expressions for the Fisher-Rao metric on the statistical manifold of the matrix Bingham parameters, and for the KL-divergence between two matrix Bingham probability densities, which allows us to efficiently compute latent geodesics. This extends the concrete benefits of latent representation geometries to subspace-based domains. We evaluate our method over both synthetic and realistic wireless communications scenarios (Sec. 5), showing its auto-encoding and latent representation capabilities. The proposed method is graphically represented in Fig. 1.

2 BACKGROUND

Variational autoencoders (VAEs, Kingma and Welling (2014)) are widely studied deep latent variable models that express a data density by marginalizing a latent variable, $p(\mathbf{x}) = \int p_\psi(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$. The *generative distribution*[1] $p_\psi(\mathbf{x}|\mathbf{z})$ is parametrized by a neural network with weights ψ, while the *prior* is usually a standard normal. The latent posterior $p(\mathbf{z}|\mathbf{x})$ is generally intractable and is approximated by an *encoder distribution* $q_\phi(\mathbf{z}|\mathbf{x})$, which is also parametrized by a neural network (with weights ϕ). The *data marginal likelihood* $p(\mathbf{x})$ is intractable, and it is lower bounded through the evidence lower bound:

$$\mathrm{L}_{\phi,\psi}(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi} \log p_\psi(\mathbf{x}|\mathbf{z}) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \tag{1}$$

where $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ is the KL-divergence. Conceptually, the VAE aims to model high-dimensional data $\mathbf{x} \in \mathcal{X}$ through a low-dimensional latent representation $\mathbf{z} \in \mathcal{Z}$. This allows for interpreting the data-generating mechanism through the latent representations.

Latent representation geometries (Arvanitidis et al., 2018; Tosi et al., 2014) support the interpretation of latent variables by providing strict identifiability guarantees (Syrota et al., 2025). These techniques assume $p_\psi(\mathbf{x}|\mathbf{z})$ to be Gaussian, and use differential geometry to bring the metric from $\mathcal{X}$ into $\mathcal{Z}$ by locally linearizing the neural network behind $p_\psi(\mathbf{x}|\mathbf{z})$. This endows the latent representation space with a Riemannian metric under which many computations have identifiable outcomes (Syrota et al., 2025).

Arvanitidis et al. (2022) recently extended these approaches to general information geometries (Amari, 2016; Ay et al., 2017) allowing them to work with non-Gaussian generative distributions $p_\psi(\mathbf{x}|\mathbf{z})$. Letting $\mathcal{H}$ denote the parameter space of the generative distribution, we can view this distribution as determined by an immersion h mapping from the latent space $\mathcal{Z}$ to the parameter space $\mathcal{H}$. The latter is naturally endowed with a Fisher-Rao metric,

$$\mathbf{I}_{\mathcal{H}}(\theta) = \int_{\mathcal{X}} (\nabla_\theta \log p(\mathbf{x}|\theta) \nabla_\theta \log p(\mathbf{x}|\theta)^\intercal) p(\mathbf{x}|\theta) \mathrm{d}\mathbf{x}. \tag{2}$$

One can show that under this Riemannian metric, changes in distribution parameters are measured infinitesimally using the KL divergence (Amari, 2016). In the context of latent variable models, this Fisher-Rao metric can be brought into the latent space as (Arvanitidis et al., 2022)

$$\mathbf{G}(\mathbf{z}) = \mathbf{J}_h^\intercal(\mathbf{z}) \mathbf{I}_{\mathcal{H}}(h(\mathbf{z})) \mathbf{J}_h(\mathbf{z}), \tag{3}$$

where $\mathbf{J}_h$ denotes the Jacobian of h evaluated at $\mathbf{z}$. Most computations performed under this Riemannian metric are identifiable (Syrota et al., 2025).

A key construction in this framework is the notion of a *shortest path* or *geodesic*. These are latent curves $c : [0, 1] \to \mathcal{Z}$ with minimal length,

$$\begin{aligned} \mathrm{L}(c) &= \int_0^1 \sqrt{\dot{h}(c(t))^\intercal \mathbf{I}_{\mathcal{H}}(h(c(t))) \dot{h}(c(t))} \mathrm{d}t \\ &= \int_0^1 \sqrt{\dot{c}(t)^\intercal \mathbf{G}(\mathbf{z}) \dot{c}(t)} \mathrm{d}t. \end{aligned} \tag{4}$$

Arvanitidis et al. (2022) noted that, since the Fisher-Rao metric locally coincides with the KL-divergence between two points of the output statistical manifold, the length of the latent curve c can be written as:

$$\mathrm{L}(c) = \lim_{n\to\infty} \sum_{n=1}^{N-1} \sqrt{D_{KL}(p(\mathbf{x}|c(t_n)) \parallel p(\mathbf{x}|c(t_{n+1})))}\, \mathrm{d}t, \tag{5}$$

which leads to an efficient approximation by truncating the sum. Geodesics can, thus, be efficiently computed

[1] We refer to $p_\psi(\mathbf{x}|\mathbf{z})$ either as *generative distribution* or as *probabilistic decoder* depending on the context.

whenever such KL-divergence can be optimized. One of our contributions is an efficient evaluation of the KL divergence between distributions over subspace-valued random variables.

The space of subspaces of given dimension is commonly denoted as the *Grassmann manifold* (Grassmann (1896)), and it plays a key role in the motivating example of Sec. 1. Consider a d-dimensional vector space V, which we will identify with $\mathbb{R}^d$. For $0 < r \leq d$, the set of r-dimensional linear subspaces of $\mathbb{R}^d$ is a Riemannian manifold.

Definition 2.1. Let V be a d-dimensional vector space over a field K and $0 < r \leq d$. The Grassmann manifold $\mathrm{Gr}_r(V)$ is the set of r-dimensional linear subspaces of V.

We denote $\mathrm{Gr}_r(\mathbb{R}^d)$ by $\mathrm{Gr}(d, r)$. There exists a bijection between the Grassmann manifold and the set of orthogonal projection matrices in $\mathbb{R}^{d\times d}$ of rank r, i.e., the set of symmetric, idempotent matrices of rank r in $\mathbb{R}^{d\times d}$ (see, e.g., Bendokat et al. (2024)):

$$\mathrm{Gr}(d, r) \cong \{P \in \mathbb{R}^{d\times d} | P^\intercal = P,\, P^2 = P,\, \mathrm{rank}(P) = r\}. \tag{6}$$

This matrix representation for points on the Grassmann manifold is commonly used computationally (Helmke et al., 2007; Huang et al., 2018).

An r-dimensional subspace $\mathcal{W}$ can be represented by any orthonormal matrix in $\mathbb{R}^{d\times r}$ whose columns span $\mathcal{W}$. The set of orthonormal matrices in $\mathbb{R}^{d\times r}$ also constitutes a Riemannian manifold named *Stiefel manifold* (Stiefel (1935)).

Definition 2.2. The Stiefel manifold $\mathrm{St}(d, r)$ consists of the set of r-frames in $\mathbb{R}^d$, i.e., orthonormal matrices in $\mathbb{R}^{d\times r}$:

$$\mathrm{St}(d, r) := \{X \in \mathbb{R}^{d\times r} | X^\intercal X = \mathbb{I}_r\}. \tag{7}$$

Since all the orthonormal bases in $\mathbb{R}^{d\times r}$ that are related by a right-orthogonal transformation span the same subspace, the Grassmann manifold can be seen as a quotient space on the Stiefel manifold under the action of the orthogonal group (Edelman et al. (1998)), i.e., $\mathrm{Gr}(d, r) = \mathrm{St}(d, r)/O(r)$, whose equivalence classes are:

$$[X] = \{XQ \,|\, Q \in O(r)\}, \tag{8}$$

with $O(r) = \{Q \in \mathbb{R}^{r\times r} | Q^\intercal Q = QQ^\intercal = \mathbb{I}_r\}$.

3 VARIATIONALLY AUTOENCODING GRASSMANN

Our first objective is to construct a deep latent variable model that emits subspaces. Technically, we target a density $p(\mathcal{W})$ for $\mathcal{W} \in \mathrm{Gr}(d, r)$, which is parametrized by a latent variable $\mathbf{z}$. Within the VAE-family of models, this requires a computationally tractable likelihood $p(\mathcal{W}|\mathbf{z})$, where we will consider the *matrix Bingham distribution* (Bingham, 1974) owing to its invariance properties and its capability to consistently model low-rank settings. We will report also the complex-valued extension of the matrix Bingham distribution as we will show its utility in modeling the wireless communications scenarios targeted in our experiments (Sec. 5).

3.1 The matrix Bingham distribution

Several probability distributions supported on the Stiefel manifold exist, e.g., the matrix Langevin distribution (also referred to as matrix von Mises-Fisher) and the matrix Bingham distribution (Chikuse, 2012). Among them, the matrix Bingham distribution (Bingham (1974)) only depends on the orthogonal projector $P = XX^\intercal$, i.e., on the subspace spanned by the random matrix $X \in \mathrm{St}(d, r)$ and not just on X. This yields useful invariance properties for the matrix Bingham distribution, which can be leveraged for uncertainty modeling over the Grassmann manifold.

The Bingham distribution was originally introduced for the hypersphere by Bingham (1974) and later studied on the Stiefel manifold by Chikuse (2012), who referred to it as the *matrix Bingham distribution*. Its probability density function is

$$p(X|M, C) = b(C)^{-1}\mathrm{etr}(CM^\intercal XX^\intercal M) \tag{9}$$

where $\mathrm{etr}(\cdot) = \exp(\mathrm{tr}(\cdot))$ is the exponential of the trace of a matrix, $X \in \mathrm{St}(d, r)$, $M \in O(d)$ is an orthogonal orientation matrix, C is a diagonal concentration matrix, and $b(C)$ is the density normalizing constant, a confluent hypergeometric function of matrix argument:

$$\begin{aligned} b(C) &= \int_{X\in\mathrm{St}(d,r)} \mathrm{etr}(CM^\intercal XX^\intercal M)\,\mathrm{d}X \\ &= {}_1F_1\left(\frac{1}{2}r; \frac{1}{2}d; C\right). \end{aligned} \tag{10}$$

Besides the standard definition, we consider also the case where some of the diagonal entries of C can be zero, so that, by the rules of the trace, the resulting M matrix, after product simplifications, is a d-dimensional k-frame $M \in \mathrm{St}(d, k)$ and $C \in \mathbb{R}^{k\times k}$. Therefore, we generalize the analysis and the derivations to this condition. The normalizing constant does not depend on the orientation matrix M and is therefore indicated as a function of C (Bingham, 1974).

The matrix Bingham density (9) only depends on X by the orthogonal projection matrix $P = XX^\intercal$. Since X is an orthonormal r-frame as $X \in \mathrm{St}(d, r)$, P represents the orthogonal projection matrix within the

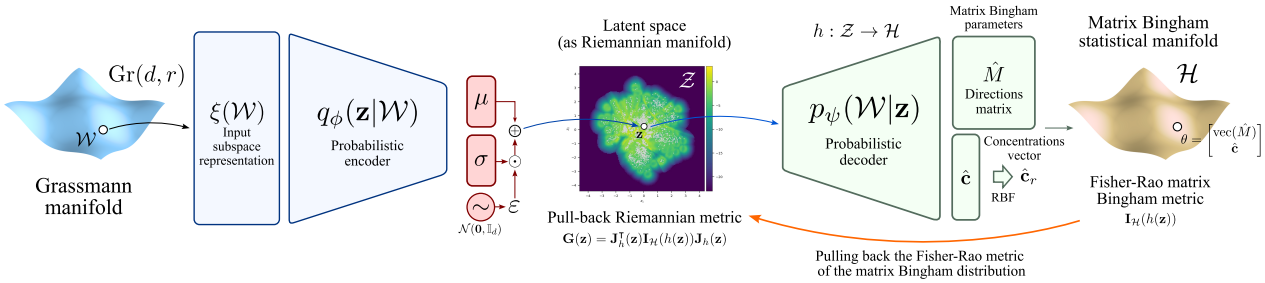


Figure 2: Proposed method. A linear subspace lying on a Grassmann manifold $\mathrm{Gr}(d, r)$ is mapped into an input matrix representation on $\mathrm{St}(d, r)$ and processed through a VAE architecture. The probabilistic output is defined over the parameter space of a matrix Bingham distribution. The latter is endowed with the Fisher-Rao metric, which provides it with the Riemannian structure of a statistical manifold. Finally, the latent space $\mathcal{Z}$ is endowed with a Riemannian metric by pulling back the Fisher-Rao metric from $\mathcal{H}$ through the smooth decoder $h : \mathcal{Z} \to \mathcal{H}$.

linear subspace spanned by the columns of X and is a matrix representative for a point in the Grassmann manifold by the isomorphism defined in (6). Moreover, the matrix Bingham distribution is invariant under right-orthogonal transformations.

Proposition 3.1. *The matrix Bingham distribution is invariant under right actions of the orthogonal group. For all $X \in St(d, r)$ and $Q \in O(r)$:*

$$p(X|M, C) = p(XQ|M, C) \tag{11}$$

Proof. See Appendix B.1. □

Whereas on the Stiefel manifold the matrix Bingham multiple modes are associated to all the orthonormal matrices spanning the same subspace, on the Grassmann manifold a single mode, identified by an r-dimensional subspace, is associated to all of them. For a matrix Bingham distribution, a mode can be easily derived as the leading eigenvectors of matrix M.

Proposition 3.2. *Let $p_{\mathcal{B}}(X|M, C)$ be a matrix Bingham distribution as in (9), with $X \in St(d, r)$, $M \in O(d)$ and $C = diag(\mathbf{c}), \mathbf{c} \in \mathbb{R}^d_+$. A mode of the matrix Bingham is given by the leading eigenvectors of $G = MCM^\intercal$.*

Proof. See Appendix I. □

Therefore, rearranging (9), we write the matrix Bingham probability density function across subspaces as:

$$p(\mathcal{W}|M, C) = b(C)^{-1}\mathrm{etr}(CM^\intercal PM), \tag{12}$$

where P is the orthogonal projection matrix defined above into the subspace $\mathcal{W}$. In the following, we will use the matrix Bingham distribution based on the Stiefel manifold, as it provides a suitable numerical representation, switching to (12) when required.

Owing to the computational complexity of evaluating a confluent hypergeometric function, to compute the matrix Bingham normalizing constant we use the approximation of ${}_pF_q$ as a truncated series of Jack functions proposed in Koev and Edelman (2006):

$$\begin{aligned} &{}^m_p F^{(\alpha)}_q(a_1, \ldots, a_p; b_1, \ldots, b_p; X) = \\ &\sum_{k=0}^{m} \sum_{\kappa \vdash k} \frac{(a_1)^{(\alpha)}_\kappa \cdots (a_p)^{(\alpha)}_\kappa}{k!(b_1)^{(a)}_\kappa \cdots (b_q)^{(\alpha)}_\kappa} \cdot C^{(\alpha)}_\kappa(X), \end{aligned} \tag{13}$$

where $p, q \geq 0$ are integers, $\alpha > 0$ is a parameter, $X \in \mathbb{R}^{n \times n}$ is a symmetric matrix, and the truncation is computed for all partitions $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_\ell)$ of k s.t. $|\kappa| \leq m$, $|\kappa| \coloneqq \kappa_1 + \kappa_2 + \cdots + \kappa_\ell$, and $\kappa_i \geq 0$ for $i \in \{1, \ldots, \ell\}$; $C^{(\alpha)}_\kappa(X)$ is the Jack function and $(a)^{(\alpha)}_\kappa$ is the generalized Pochhammer symbol.

Besides approximating the value of the normalizing constant, this also allows for differentiation with respect to the diagonal values of the concentration matrix C. We will show the utility of this in the evaluation of the KL-divergence between two matrix Bingham distributions proposed in Sec. 4.2. We refer the reader to Appendix C, where we provide more details on the approximation and we discuss the impact of the truncation error on the computation of the latent geodesics and on the definition of the matrix Bingham likelihood.

As indicated by Kent (1994) and Bingham et al. (1992) for the hypersphere case, the complex matrix Bingham distribution can be represented by a real-valued matrix Bingham distribution over a twice dimensional space. We introduce this result here as it will be relevant in Sec. 5.2 on the wireless communications experiments, which involve matrix representations that are intrinsically complex-valued and can benefit from this formalization.

Proposition 3.3. *Let $\tilde{X} \in St_{\mathbb{C}}(d, r)$ and $X \in St(2d, 2r)$. The complex-valued matrix Bingham distribution $p_{\mathcal{CB}}(\tilde{X}|\tilde{M}, \tilde{C})$ is equivalent to a real-valued*

matrix Bingham distribution $p_{\mathcal{B}}(X|M,C)$ *s.t.*

$$M = \begin{bmatrix} \tilde{M}_r & -\tilde{M}_i \\ \tilde{M}_i & \tilde{M}_r \end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \quad C = \begin{bmatrix} \tilde{C} & \mathbf{0} \\ \mathbf{0} & \tilde{C} \end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \quad X = \begin{bmatrix} \tilde{X}_r & -\tilde{X}_i \\ \tilde{X}_i & \tilde{X}_r \end{bmatrix} \in \mathbb{R}^{2d\times 2r}, \tag{14}$$

where $A_r = \mathrm{Re}(A)$ *and* $A_i = \mathrm{Im}(A)$ *denote, respectively, the real and imaginary parts of a matrix* A*.*

Proof. See Appendix J. □

The reformulation of the complex-valued matrix Bingham likelihood into a real-valued one of doubled dimension is purely algebraic and relies on the representation of complex-valued matrix multiplication as an operation between suitably structured real-valued matrices.

Leveraging the methods discussed in Sec. 2, we aim to learn an uncertainty-aware smooth immersion $h : \mathcal{Z} \to \mathcal{H}$ charting a Riemannian manifold $\mathcal{M} = h(\mathcal{Z})$ in the parameter space of the matrix Bingham distribution. In Fig. 2, we graphically depict this relationship.

3.2 Group-invariant subspace encoding

After the definition of the properties of the probabilistic decoder for uncertainty modeling on $\mathrm{Gr}(d,r)$, we focus here on the representation of the VAE input on the Grassmann manifold. Since matrices $X \in \mathrm{St}(d,r)$ related by a right orthogonal transformation identify the same subspace, the mapping between r-frames and r-dimensional subspaces is not injective.

Our aim is to obtain an encoding architecture that suitably represents the subspaces while being invariant with respect to a chosen representative basis for the subspace. Two such candidates are the orthogonal projection matrix input representation and an encoding architecture invariant to right-orthogonal transformations of an input orthonormal basis. We discuss both options in detail in Appendix L.

In our experiments, we employ an $O(r)$-invariant VAE encoding architecture derived from Huang et al. (2018). We consider as input an arbitrary basis $X \in \mathrm{St}(d,r)$ for a subspace $\mathcal{W} \in \mathrm{Gr}(d,r)$. We notice that an input representation on $\mathrm{St}(d,r)$ has a lower dimensionality with respect to a projection matrix.

4 GRASSMANN LATENT INFORMATION GEOMETRY

Having established suitable machinery for building VAEs over the Grassmann manifold, we next target the information geometry associated with the latent variables of the model. For this, we need tractable expressions for the Fisher-Rao metric associated with the matrix Bingham distribution, and ditto for the KL-divergence of matrix Bingham distributed variables. We also need to ensure the differentiability of the approximation of the matrix Bingham normalizing constant in (13) to model the output likelihood to use it during model training and for the computation of the KL-divergence and the Fisher-Rao metric. We derive these expressions as novel contributions of this paper since not already present in the literature to the best of our knowledge.

4.1 The Fisher-Rao metric for the matrix Bingham statistical manifold

We investigate here the Fisher-Rao metric for the statistical manifold of the matrix Bingham distribution parameters. Referring to (9), let $\mathbf{m} = \mathrm{vec}(M) \in \mathbb{R}^{d\cdot k}$ be the vectorization of the matrix Bingham direction matrix M, and let $\mathbf{c} = \mathrm{diag}(C) \in \mathbb{R}^k$. We consider the parameter space $\mathcal{H}$ defined by the parameters $\theta = [\mathbf{m}, \mathbf{c}]$. In order to define the metric associated to the statistical manifold, our aim is to determine the Fisher information matrix using the parameters vector θ.

Under the required regularity conditions, we write the Fisher information matrix in terms of the expected value of the Hessian of the matrix Bingham log-likelihood. The parameters M and C are considered in their respective vector forms $\mathbf{m}$ and $\mathbf{c}$ to simplify the representation of the information matrix, which would result in a 4th-order tensor in the matrix case.

For the matrix Bingham case, the Hessian of the log-likelihood function can be derived from (9) by applying the matrix derivation rules and vectorizing the results. We distinguish three cases to differentiate the log-likelihood with respect to M and to C, according to the differentiation order and taking into consideration that $\mathbf{I}_{\mathcal{H}}$ is symmetric.

Proposition 4.1. *The Fisher information matrix* $\boldsymbol{I}_{\mathcal{H}}$ *of a matrix Bingham random variable* $X \sim p(X|M,C)$*, with* $M \in St(d,k)$ *and* $C \in \mathbb{R}^{k\times k}$ *diagonal, can be written as the block matrix:*

$$\boldsymbol{I}_{\mathcal{H}}(\theta) = \begin{bmatrix} \boldsymbol{I}_{\mathcal{H},\mathbf{mm}} & \boldsymbol{I}_{\mathcal{H},\mathbf{mc}} \\ \boldsymbol{I}_{\mathcal{H},\mathbf{cm}} & \boldsymbol{I}_{\mathcal{H},\mathbf{cc}} \end{bmatrix}, \tag{15}$$

with

$$[\boldsymbol{I}_{\mathcal{H},\mathbf{mm}}]_{(ij)(m\ell)} = -\mathbb{E}_p[Hessf(M)]_{(ij)(m\ell)}$$

$$[\boldsymbol{I}_{\mathcal{H},\mathbf{cc}}]_{ij} = \frac{1}{\bar{b}(\boldsymbol{c})}\frac{\partial^2 \bar{b}(\boldsymbol{c})}{\partial \boldsymbol{c}_i \partial \boldsymbol{c}_j} - \left[\frac{1}{\bar{b}(\boldsymbol{c})^2}(\nabla \bar{b}(\boldsymbol{c}))^\intercal \nabla \bar{b}(\boldsymbol{c})\right]_{ij}$$

$$[\boldsymbol{I}_{\mathcal{H},\mathbf{mc}}]_{(ij)\ell} = -\mathbb{E}_p(\delta_{\ell j}[MM^T PM]_{ij} + 2\delta_{j\ell}[PM]_{ij} + M_{i\ell}[M^T PM]_{\ell j}),$$

where $f(M) = p(X|M, C)$, δ_{ij} is the Kronecker delta, $\mathbf{c}$ is the diagonal of matrix C, $\bar{b}(\boldsymbol{c}) = b(diag(\boldsymbol{c}))$, $\nabla\bar{b}(\boldsymbol{c})$ is considered as a row vector, $(ij) = (j-1)\cdot d + i$ is the vectorized index associated to the element M_{ij} of matrix $M \in \mathbb{R}^{d\times k}$, $P = \mathbb{E}_p[XX^\intercal]$, and Hessf denotes the Riemannian Hessian of f.

Proof. Proof and further details in Appendix D. □

In Sec. 5, we use the derived Fisher information matrix to define the metric for the matrix Bingham statistical manifold and we pull it back to the VAE latent space according to (23). This defines the geometry of the latent space through the learned probabilistic decoder and the ambient space structure given by the information geometry of the Grassmann manifold.

4.2 The KL-divergence between matrix Bingham distributions

Our aim is now to determine an efficient procedure to compute the KL-divergence between two matrix Bingham distributions in order to access fast geodesics computations using the method proposed in Arvanitidis et al. (2022), which leverages automatic differentiation.

Expanding on the analysis proposed in Kurz et al. (2014) for hyperspherical Bingham distributions, we derive the KL-divergence between two matrix Bingham distributed variables. We show that the KL-divergence between two matrix Bingham distributions is available in closed form. This derivation is novel for the matrix Bingham distribution to the best of our knowledge.

Proposition 4.2. *The KL-divergence between two matrix Bingham distributions P and Q with probability density functions $p(X|M_1, C_1)$ and $q(X|M_2, C_2)$, respectively, is*

$$D_{\mathrm{KL}}(p \parallel q) = \int_{St(d,r)} p(X|M_1, C_1) \log\left(\frac{p(X|M_1, C_1)}{q(X|M_2, C_2)}\right) \mathrm{d}X$$
$$= \log\frac{b(C_2)}{b(C_1)} + Tr\left((M_1 C_1 M_1^\intercal - M_2 C_2 M_2^\intercal)\mathbb{E}_p[XX^\intercal]\right), \tag{16}$$

where the integral is performed over the Stiefel manifold $St(d, r)$ and the normalizing constants $b(C_1)$ and $b(C_2)$ are defined as in (10).

Proof. See Appendix E. □

The evaluation of the KL-divergence requires the computation of the expected value of $XX^\intercal$ with respect to $p(X|M_1, C_1)$. We provide an explicit formula for this term depending only on the distribution parameters.

Proposition 4.3. *The expected value of $XX^\intercal$ with respect to $p(X|M, C)$ is:*

$$\mathbb{E}_p[XX^\intercal] = \frac{1}{b(C)} M\, diag\left(\frac{\partial b(C)}{\partial \mathbf{c}_1}, \cdots, \frac{\partial b(C)}{\partial \mathbf{c}_d}\right) M^\intercal \tag{17}$$

Proof. See Appendix F for the derivation of $\mathbb{E}_p[XX^\intercal]$ and Appendix H for a discussion on the differentiation of $b(C)$ with respect to the diagonal elements of C. □

As for the likelihood (9), the efficiency of the KL-divergence computation strictly depends on the computational complexity of the algorithm used to evaluate the matrix Bingham normalizing constant. Also in this case, we exploit the approximation proposed by Koev and Edelman (2006) and discussed in detail in Appendix C to efficiently evaluate the normalizing constant of the matrix Bingham in (16) and its gradient with respect to the diagonal elements of C in (17).

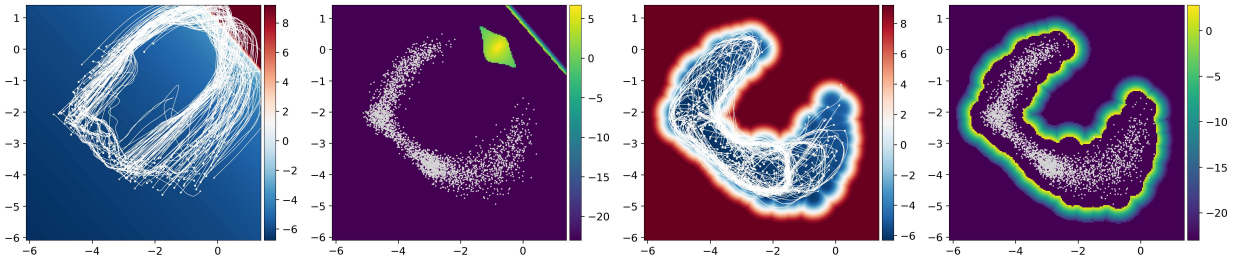
5 EXPERIMENTS

In this section, we propose a set of numerical experiments assessing the meaningfulness of the uncertainty-aware latent representations obtained in the latent space $\mathcal{Z}$. First, we discuss a synthetic experiment on the set of 1-dimensional subspaces of $\mathbb{R}^3$, i.e., $\mathrm{Gr}(3, 1)$. This can be useful for visualization in a low-dimensional output space (Sec. 5.1). Then, we explore the model capabilities on the representation of the space-time features of high-frequency EM propagation environments (Sec. 5.2) for reliable wireless communications, i.e., the real-world problem motivating this work.

For each experiment, we trained a VAE with the model architecture depicted in Fig. 2. The decoder is split into the prediction of the M directional matrix and of the diagonal of the C concentration matrix of a matrix Bingham probability density $p(X|M, C)$. M is an orthonormal matrix and is estimated by polar expansion through the singular vectors of an unstructured matrix Y having the same dimensions of M using thin-SVD, whose differentiability is proven by Townsend (2016):

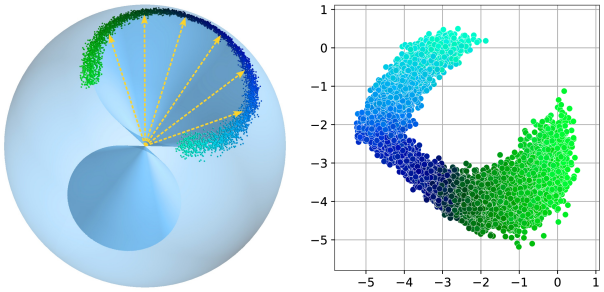
$$U, s, V^\intercal = \mathrm{SVD}_{\mathrm{th}}(Y(\mathbf{z})), \quad M = UV^\intercal. \tag{18}$$

The diagonal elements of C model the uncertainty as the inverse of the variance for the corresponding directions estimated in matrix M. M and C are learned in an alternating optimization: first, M is learned by fixing the diagonal values of the concentration matrix to a high value (i.e., 30); then, all the layers are jointly trained to learn also the concentration matrix C.



(a) Geodesics w/o u.m. (b) Pullback metric w/o u.m. (c) Geodesics with u.m. (d) Pullback metric with u.m.

Figure 3: Results on synthetic data on Gr(3, 1). a) and c) depict a set of sample geodesics computed between latent points for the models with and w/o uncertainty modeling (u.m.); the background represents the inverse of the predicted average concentration. b) and d) represent $\log(\sqrt{\det(\mathbf{G}(\mathbf{z}))})$ with and w/o uncertainty modeling.



(a) Input space (b) Latent space

Figure 4: a) Representation of the input space for the synthetic experiment. The dashed arrows highlight the C-shaped samples on the unit 2-sphere ($\cong \mathrm{St}(3,1)$). b) Latent representations obtained after model training. The color gradient shows how the input subspaces are mapped by the encoder into the latent space.

5.1 Synthetic experiment

For visualization and validation purposes, we consider a simple synthetic experiment on Gr(3, 1). When the subspaces dimension r is 1, the matrix Bingham distribution reduces to a Bingham distribution on the hypersphere (Bingham (1974)). We choose as inputs a set of unit vectors representing 1-dimensional subspaces. Therefore, the input coordinates can be equivalently represented as points on the 2-sphere. We consider a cone through the origin and we sample a set of points on the centered unit 2-sphere around its intersection with the cone to stochastically define a C-shaped data pattern, as depicted in Fig. 4a. The hyperparameters, model details and training procedure for this experiment are reported in Appendix M.

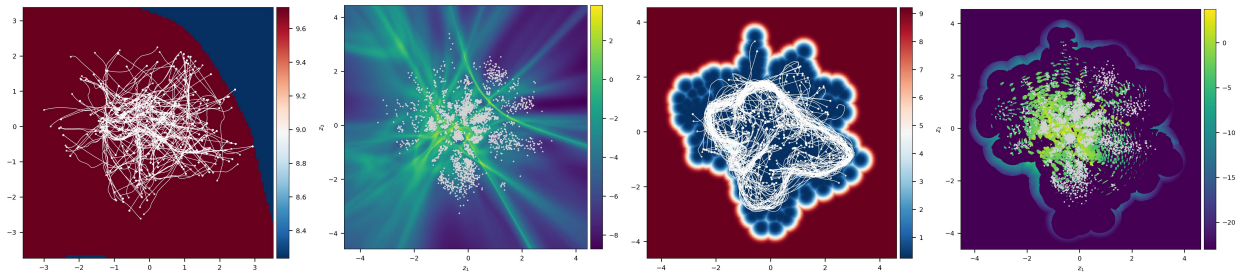
In Fig. 3, we provide the results for the synthetic experiment, where we analyzed the geodesics and the log of the pull-back metric volume for the models with and without meaningful uncertainty modeling, which has been shown in Arvanitidis et al. (2022) to be critical to capture the latent geometry. Uncertainty quantification is performed by regularizing the diagonal of the concentration matrix C outside the latent data support in (9). We refer to Appendix A for further details. As depicted in Fig. 3c, in the presence of uncertainty modeling, the geodesics are *constrained* in the neighborhood of the data support within the latent space. The pullback metric in Fig. 3d is consistent with the behavior of the obtained geodesics. By contrast, the absence of uncertainty modeling leads to unstructured geodesics (Fig. 3a) and pull-back (Fig. 3b).

In Fig. 4a, we depict the input coordinates on the 2-sphere along with the points sampled on a C-shaped data pattern. Fig. 4b shows the set of input points mapped into the 2-dimensional latent space, where the point colors match the ones within the input space.

5.2 High-freq. EM propagation experiments

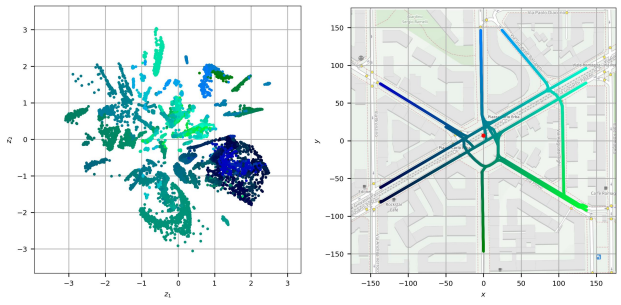
In this experiment, we investigate latent subspace modeling for high-frequency EM propagation environments. In wireless communications, the properties of the radio propagation environment (commonly referred to as *wireless channel*) between a transmitter (Tx) and a receiver (Rx) are routinely estimated. This step is fundamental to compensate for the distortions produced by the environment on the transmitted signal (Goldsmith (2005); Spagnolini (2018)). At high frequencies, the spatial and temporal features of the channel can be effectively modeled through a set of representative subspaces (Brighente et al. (2020)), which are the target of this experiment. The effective estimation and modeling of such subspaces can considerably improve the reliability and performance of a wireless communication link, which can be particularly relevant in safety-critical operational conditions or in potentially noisy and low-data scenarios like high-mobility ones.

We focus on the spatial features of the wireless channel.



(a) Geodesics w/o u.m. (b) Pullback metric w/o u.m. (c) Geodesics with u.m. (d) Pullback metric with u.m.

Figure 5: Latent space structural properties achieved on the EM propagation experiment over $\mathrm{Gr}(128, 6)$ and $\mathrm{Gr}(32, 4)$. a) and c) depict a set of sample geodesics computed between latent points for the models with and w/o uncertainty modeling (u.m.); the background represents the inverse of the predicted average concentration. b) and d) represent $\log(\sqrt{\det(\mathbf{G}(\mathbf{z}))})$ with and w/o uncertainty modeling.



(a) Latent space (mean) (b) Roads geometry

Figure 6: a) Latent representations (means) for the training data points obtained for the high-frequency EM propagation experiment. b) Top view of the considered urban vehicular scenario. The roads' 2D color gradient is mapped to the colors of the latent points.

We consider a narrow-band vehicle-to-infrastructure (V2I) wireless communication system operating at 28 GHz carrier frequency. The system is featured by a receiving base station (BS) and a set of transmitting vehicular equipment (VEs) crossing an urban setting. We simulate a realistic urban radio propagation environment and vehicular traffic where a set of spatial channel subspaces is estimated at the BS based on the communication signals received from the VEs. We provide in Appendix N the details on wireless communication system and channel models, subspace estimation and simulation framework.

We separate the spatial channel subspace components at the Tx and at the Rx, which leads to higher computational efficiency at the expense of minimal performance loss for the purposes of this experiment. The resulting estimated subspaces lie on the complex Grassmann manifolds $\mathrm{Gr}_{\mathbb{C}}(64, 3)$ and $\mathrm{Gr}_{\mathbb{C}}(16, 2)$ for the Tx and Rx spatial components, respectively. As shown in Proposition 3.3, to stochastically represent such manifolds through a matrix Bingham distribution, they can be equivalently mapped into the real-valued Grassmann manifolds $\mathrm{Gr}(128, 6)$ and $\mathrm{Gr}(32, 4)$.

To encode such representations, we devised an architecture based on Huang et al. (2018). For visualization purposes, we explore a 2-dimensional latent space and we adopt a fully-connected decoder to map the latent representations into the matrix Bingham distribution parameters. We refer to Appendix N for extensive details on the adopted VAE model architecture, selected hyperparameters, training procedure, and experiments.

Latent representation geometry. In Fig. 5, we provide the results achieved in latent space modeling of the targeted spatial channel subspaces. We show in Fig. 5c and Fig. 5d that modeling the latent Riemannian structure through an uncertainty-aware pull-back metric *constrains* the geodesics near the data support, consistently with the observations in the synthetic case. The background represents the model variance as the inverse of the predicted average concentration in the matrix Bingham parameters, showing that the model consistently predicts a low variance near the data support and a high variance far from it.

Fig. 10a shows the achieved mean latent points over the training dataset, where we obtained an analogous distribution on the validation set. The colors of the points correspond to the ones of the vehicular trajectories depicted in Fig. 6b. In the latent space, the subspaces appear to be clustered according to the proximity of the trajectory points from which the corresponding spatial subspaces were sampled.

Reconstruction capabilities. To assess the reconstruction capabilities of the VAE, we used the subspaces estimated through the matrix Bingham mode (see Prop. 3.2) to filter a set of Least Squares (LS)

channel estimates generated for the same VE locations as for the input channel subspaces. We performed low-rank (LR) channel estimation (Mizmizi et al. (2021)) using the predicted subspaces, and we measured the normalized mean squared error (NMSE) of the LS and LR estimates. We name *NMSE ratio* the ratio between the NMSE of the LR estimates and the one of the LS estimates (the lower the NMSE ratio, the better).

Our model converged to ≈ -19.7 dB training NMSE ratio between LR and LS estimation and ≈ -19.6 dB validation NMSE ratio with respect to the ground truth performance of ≈ -22 dB. This showcases the generalization capabilities of the model with respect to the channel estimates sampled over different wireless channel realizations across urban vehicular trajectories.

To achieve a more direct comparison with respect to previous methods, we have compared our model with the subspace regressor proposed in (Cazzella et al., 2022) over the same scenario. Both methods attained the ground truth reference NMSE performance by a low margin, with our method additionally providing latent geometrical modeling and uncertainty quantification over the predicted subspaces.

Evaluation across different conditions. To extensively validate the properties of the proposed model, we tackled different communication conditions. We considered a different SNR at -30 dB for the wireless communications scenario examined above. Also in this case, we obtained a -22.06 dB NMSE ratio on the validation set. We considered experiments on a set of 5 different randomization seeds for the same scenario in Fig. 6b, obtaining a mean NMSE ratio of -18.24 dB over the range (-19.6 dB, -17.61 dB), along with well-structured latent spaces. Moreover, we selected another urban scenario with different structural conditions with respect to the one introduced above, obtaining a comparable NMSE ratio of -17.4 dB on the validation set, as detailed in Appendix N.

6 RELATED WORK

A set of methods exploiting the representational capabilities of neural networks for subspace modeling has been recently proposed. Izmailov et al. (2020) have developed a subspace inference approach to face the challenges induced by high dimensionality in Bayesian deep learning models. Deep subspace encoders have instead been proposed for nonlinear system identification in (Beintema et al., 2023). Huang et al. (2018) have recently introduced GrNet, a deep learning architecture targeting the Grassmann manifold, which processes subspace input representations through a sequence of suitably designed layers towards classification tasks. Yataka et al. (2023) have proposed a generative model based on continuous normalization flows on the Grassmann manifold to effectively capture the shape information in the generative modeling of 3D shapes, while in (Cribeiro-Ramallo et al., 2025), the V-GAN model has been developed for adversarial subspace generation. In the wireless communications literature, Cazzella et al. (2022) have proposed a subspace-based non-generative neural network architecture to improve the performance and reliability of high-frequency wireless communication systems.

The proposed approach differs from the methods currently available in the literature since it aims to devise a latent variable model that operates over subspaces and produce uncertainty-aware and identifiable latent representations leveraging information geometry.

7 CONCLUSION

In this paper, we explored the latent information geometry of the Grassmann manifold. First, we devised a VAE architecture operating over subspaces lying on a Grassmann manifold, selecting the matrix Bingham distribution as a likelihood and a suitable subspace encoding architecture based on their group-invariance and computational properties. Then, we investigated the geometric structure of the latent space aiming to attain uncertainty-aware and identifiable latent representations. Leveraging latent information geometries, we showed how to derive the Fisher-Rao metric on the statistical manifold of the matrix Bingham parameters and to pull it back into the latent space. Then, we derived the KL-divergence between two matrix Bingham-distributed random variables to allow for efficient geodesics computation. We provide numerical results on both synthetic and real-world scenarios showcasing the effectiveness of the proposed method in deriving meaningful latent representations and achieving consistent auto-encoding performance.

Limitations. Among the limitations and prospective work of this paper are the study of more specific auto-encoding architectures suited for linear subspace processing, e.g., GRNet (Huang et al., 2018) and ManifoldNet (Chakraborty et al., 2020). The complexity of the matrix Bingham normalization constant approximation in (Koev and Edelman, 2006) can become intractable for high-dimensional ambient spaces or when the full concentration matrix is considered. The derivation of fast and accurate approximation methods is required to achieve a higher efficiency during the VAE training phase. We made no particular assumptions on the properties of the auto-encoded subspaces. Nevertheless, we expect this method to be considerably more useful when the provided subspaces satisfy some form of intrinsic continuity in a suitable domain, which can be exploited through the learned latent manifold.

Acknowledgements

This work was supported by a research grant (42062) from VILLUM FONDEN. This project received funding from the European Research Council (ERC) under the European Union's Horizon research and innovation programme (grant agreement 101125993). The work was partly funded by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). This paper was supported by CINECA (Italian inter-university consortium for supercomputing) under the ISCRA-C "RAIGAMDA" and "RADARAI" grants. This work was supported by the DFF Sapere Aude Starting Grant "GADL". This work was also supported by the NVIDIA Academic Grant Program under the "Physics-Informed AI-based Channel Simulation for V2X Communications" project.

References

M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip. Millimeter Wave Channel Modeling and Cellular Capacity Evaluation. *IEEE Journal on Selected Areas in Communications*, 32(6):1164–1179, 2014.

S. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models. In *International Conference on Learning Representations*, 2018.

G. Arvanitidis, M. González-Duque, A. Pouplin, D. Kalatzis, and S. Hauberg. Pulling back information geometry. In *International Conference on Artificial Intelligence and Statistics*, pages 4872–4894. PMLR, 2022.

N. Ay, J. Jost, H. Vân Lê, and L. Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

G. I. Beintema, M. Schoukens, and R. Tóth. Deep subspace encoders for nonlinear system identification. *Automatica*, 156:111210, 2023.

T. Bendokat, R. Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1):1–51, 2024.

C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.

C. Bingham, T. Chang, and D. Richards. Approximating the matrix Fisher and Bingham distributions: Applications to spherical regression and Procrustes analysis. *Journal of Multivariate Analysis*, 41(2): 314–337, 1992.

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

A. Brighente, M. Cerutti, M. Nicoli, S. Tomasin, and U. Spagnolini. Estimation of Wideband Dynamic mmWave and THz Channels for 5G Systems and Beyond. *IEEE Journal on Selected Areas in Communications*, 38(9):2026–2040, 2020.

R. W. Butler and A. T. Wood. Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30(4):1155–1177, 2002.

L. Cazzella, D. Tagliaferri, M. Mizmizi, D. Badini, C. Mazzucco, M. Matteucci, and U. Spagnolini. Deep learning of transferable MIMO channel modes for 6G V2X communications. *IEEE Transactions on Antennas and Propagation*, 70(6):4127–4139, 2022.

R. Chakraborty, J. Bouza, J. H. Manton, and B. C. Vemuri. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2020.

Y. Chikuse. The matrix angular central gaussian distribution. *Journal of Multivariate Analysis*, 33(2): 265–274, 1990.

Y. Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2012.

J. Cribeiro-Ramallo, F. Matteucci, P. Enciu, A. Jenke, V. Arzamasov, T. Strufe, and K. Böhm. Adversarial subspace generation for outlier detection in high-dimensional data. *arXiv preprint arXiv:2504.07522*, 2025.

N. S. Detlefsen, A. Pouplin, C. W. Feldager, C. Geng, D. Kalatzis, H. Hauschultz, M. González-Duque, F. Warburg, M. Miani, and S. Hauberg. StochMan. *GitHub*, 2021.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

D. Eklund and S. Hauberg. Expected path length on random manifolds. *arXiv preprint arXiv:1908.07377*, 2019.

I. Gilitschenski, G. Kurz, S. J. Julier, and U. D. Hanebeck. Unscented orientation estimation based on the Bingham distribution. *IEEE Transactions on Automatic Control*, 61(1):172–177, 2015.

A. Goldsmith. *Wireless communications*. Cambridge university press, 2005.

H. Grassmann. *Die Ausdehnungslehre*. TCF Enslin (A. Enslin), 1896.

S. Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018.

U. Helmke, K. Hüper, and J. Trumpf. Newton's method on Grassmann manifolds. *arXiv preprint arXiv:0709.2205*, 2007.

J. Hoydis, F. A. Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller. Sionna RT: Differentiable ray tracing for radio propagation modeling. *arXiv preprint arXiv:2303.11103*, 2023.

Z. Huang, J. Wu, and L. Van Gool. Building deep networks on grassmann manifolds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020.

J. T. Kent. The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(2):285–299, 1994.

D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

P. Koev and A. Edelman. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):833–846, 2006.

A. Kume, S. P. Preston, and A. T. Wood. Saddlepoint approximations for the normalizing constant of Fisher–Bingham distributions on products of spheres and Stiefel manifolds. *Biometrika*, 100(4):971–984, 2013.

G. Kurz, I. Gilitschenski, S. Julier, and U. D. Hanebeck. Recursive Bingham filter for directional estimation involving 180 degree symmetry. *Journal of Advances in Information Fusion*, 9(2):90–105, 2014.

P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

M. Mizmizi, D. Tagliaferri, D. Badini, C. Mazzucco, and U. Spagnolini. Channel estimation for 6g v2x hybrid systems using multi-vehicular learning. *IEEE Access*, 9:95775–95790, 2021.

D. A. Roberts and L. R. Roberts. QR and LQ Decomposition Matrix Backpropagation Algorithms for Square, Wide, and Deep–Real or Complex–Matrices and Their Software Implementation. *arXiv preprint arXiv:2009.10071*, 2020.

N. Skafte, M. Jørgensen, and S. Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

U. Spagnolini. *Statistical signal processing in engineering*. John Wiley & Sons, 2018.

E. Stiefel. *Richtungsfelder und Fernparallelismus in n-dimensionalen Mannigfaltigkeiten*. PhD thesis, ETH Zurich, 1935.

S. Syrota, Y. Zainchkovskyy, J. Xi, B. Bloem-Reddy, and S. Hauberg. Identifying metric structures of deep latent variable models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research (PMLR), 2025.

A. Tosi, S. Hauberg, A. Vellido, and N. D. Lawrence. Metrics for Probabilistic Geometries. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, Quebec, Canada, July 2014.

J. Townsend. Differentiating the singular value decomposition. Technical report, Technical Report 2016, https://j-towns.github.io/papers/svd-derivative, 2016.

R. Yataka, K. Hirashima, and M. Shiraishi. Grassmann manifold flows for stable shape generation. *Advances in Neural Information Processing Systems*, 36:72377–72411, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials: On the Latent Information Geometry of the Grassmann Manifold

A THE LATENT GEOMETRY OF DEEP GENERATIVE MODELS

Latent representation geometries have been introduced in (Arvanitidis et al., 2018; Tosi et al., 2014) to support the interpretation of latent variables by providing strict identifiability guarantees (Syrota et al., 2025). In these methods, $p_\psi(\mathbf{x}|\mathbf{z})$ is assumed to be Gaussian and the latent space is endowed with the pull-back metric brought from the observation space $\mathcal{X}$. This endows the latent representation space with a Riemannian metric under which many computations have identifiable outcomes (Syrota et al., 2025).

Under mild architectural conditions, i.e., the decoder is smooth and is an immersion between $\mathcal{Z}$ and $\mathcal{X}$, the VAEs with Gaussian decoders can be endowed with a stochastic geometry representing the training data (Arvanitidis et al. (2018)). Reframing a Gaussian stochastic generator $f(\mathbf{z}) = \mu(\mathbf{z}) + \sigma(\mathbf{z}) \odot \varepsilon$ as a random projection of a deterministic manifold (Eklund and Hauberg, 2019), we obtain

$$f(\mathbf{z}) = [\mathbb{I}_D, \mathrm{diag}(\varepsilon)] \begin{bmatrix} \mu(\mathbf{z}) \\ \sigma(\mathbf{z}) \end{bmatrix} = \mathbf{P}\, \phi(\mathbf{z}), \tag{19}$$

where $\mathbf{P}$ can be interpreted as a random projection matrix and $\phi(\mathbf{z})$ encloses the distribution parameters.

We can endow the VAE latent space with a Riemannian manifold structure whose metric

$$\mathbf{G}(\mathbf{z}) = \mathbf{J}_\mu(\mathbf{z})^\intercal \mathbf{J}_\mu(\mathbf{z}) + \mathbf{J}_\sigma(\mathbf{z})^\intercal \mathbf{J}_\sigma(\mathbf{z}), \tag{20}$$

is derived by pulling back the observation space Euclidean metric through the trained decoder. Here, $\mu : \mathcal{Z} \to \mathcal{X}$ and $\sigma : \mathcal{Z} \to \mathbb{R}^D_>$ are neural networks parameterizing mean and standard deviation of the output Gaussian, and $\mathbf{J}_\mu$ and $\mathbf{J}_\sigma$ are the Jacobians of μ and σ, respectively.

From the definition of the metric, the length of a curve $c : [0, 1] \to \mathcal{Z}$ in the latent space is:

$$L(c) = \int_0^1 \|\dot{f}(c(t))\| \mathrm{d}t = \int_0^1 \sqrt{\dot{c}(t)^\intercal \mathbf{G}(c(t)) \dot{c}(t)} \mathrm{d}t. \tag{21}$$

This approach has been recently extended in Arvanitidis et al. (2022) to general information geometries (Amari, 2016; Ay et al., 2017). This extension is instrumental to work with non-Gaussian generative distributions $p_\psi(\mathbf{x}|\mathbf{z})$. We denote with $\mathcal{H}$ the parameter space of the generative distribution. In these models, the generator can be seen as an immersion $h : \mathcal{Z} \to \mathcal{H}$ that maps the latent space into the parameter space.

The latter is naturally endowed with the Fisher-Rao metric (Amari, 2016)

$$\mathbf{I}_\mathcal{H}(\theta) = \int_\mathcal{X} (\nabla_\theta \log p(\mathbf{x}|\theta) \nabla_\theta \log p(\mathbf{x}|\theta)^\intercal) p(\mathbf{x}|\theta) \mathrm{d}\mathbf{x}. \tag{22}$$

It can be shown that, under this Riemannian metric, changes in distribution parameters are measured infinitesimally using the KL divergence (Amari, 2016). In latent variable models, this Fisher-Rao metric can be brought into the latent space as (Arvanitidis et al., 2022)

$$\mathbf{G}(\mathbf{z}) = \mathbf{J}_h^\intercal(\mathbf{z}) \mathbf{I}_\mathcal{H}(h(\mathbf{z})) \mathbf{J}_h(\mathbf{z}), \tag{23}$$

where $\mathbf{J}_h$ denotes the Jacobian of h evaluated at $\mathbf{z}$. Most computations performed under this Riemannian metric are identifiable (Syrota et al., 2025).

In this framework, *latent geodesics* represent a key notion. These are latent curves $c : [0, 1] \to \mathcal{Z}$ with minimal length, defined as

$$\begin{aligned} \mathrm{L}(c) &= \int_0^1 \sqrt{\dot{h}(c(t))^\intercal \mathbf{I}_{\mathcal{H}}(h(c(t)))\dot{h}(c(t))}\mathrm{d}t \\ &= \int_0^1 \sqrt{\dot{c}(t)^\intercal \mathbf{G}(\mathbf{z})\dot{c}(t)}\mathrm{d}t. \end{aligned} \tag{24}$$

Since the Fisher-Rao metric locally coincides with the KL-divergence between two points of the output statistical manifold, the length of the latent curve c can be written as (Arvanitidis et al., 2022):

$$\mathrm{L}(c) = \lim_{n\to\infty} \sum_{n=1}^{N-1} \sqrt{D_{KL}(p(\mathbf{x}|c(t_n)) \parallel p(\mathbf{x}|c(t_{n+1})))}\,\mathrm{d}t, \tag{25}$$

which leads to an efficient approximation by truncating the sum. Geodesics can, thus, be efficiently computed leveraging such KL-divergence optimization. We show how to compute the KL-divergence between two matrix Bingham distributed random variables in Appendix E.

A.1 Uncertainty quantification and latent representations geometries

Uncertainty regularization during model training is fundamental for consistent uncertainty quantification and to learn useful latent representations geometries (Hauberg, 2018). We can regularize the model uncertainty by extrapolating to higher uncertainty values when a latent point is far from the latent encodings of the training data. As suggested in Arvanitidis et al. (2022) for consistent uncertainty modeling around the support of the data mapped to the latent space, the output concentration values in C are extrapolated to 0 when the predicted latent codes are distant from the training latent codes. To perform this step, the training latent codes are clustered using the K-means algorithm producing k training cluster centers $\{\mathbf{a}_j\}_{j=1}^k$; then, for a predicted latent code $\mathbf{z}$, each output diagonal value in $C(\mathbf{z})$ is extrapolated using $\sigma_\beta(\min_j\{\|\mathbf{z} - \mathbf{a}_j\|^2\})$, where $\sigma_\beta(d)$ is the modified sigmoid function

$$\sigma_\beta(d) = \sigma\left(\frac{d - c \cdot \mathrm{Softplus}(\beta)}{\mathrm{Softplus}(\beta)}\right),$$

proposed in Skafte et al. (2019), where σ is the sigmoid function and β and c are hyperparameters controlling the properties of uncertainty quantification.

B PROPERTIES OF THE MATRIX BINGHAM DISTRIBUTION

In this section, we discuss the properties of the matrix Bingham distribution that are relevant for the methods discussed in this paper. First, we will discuss its invariance with respect to right orthogonal transformations of the random variable; then, we discuss two implications of this property, i.e., the multimodality of the matrix Bingham distribution and its antipodal symmetry.

B.1 Invariance with respect to right orthogonal transformations

One of the main properties by which we consider the Bingham distribution as representative for the uncertainty of subspaces is its invariance with respect to right multiplications of arbitrary orthogonal matrices.

Proposition B.1. *The matrix Bingham distribution is invariant under right actions of the orthogonal group. For all $X \in St(d, r)$ and $Q \in O(r)$:*

$$p(X|M, C) = p(XQ|M, C). \tag{26}$$

Proof. Let $X \in \mathrm{St}(d, r)$ and $X' = XQ$, with $Q \in \mathcal{O}(r)$ an arbitrary orthogonal matrix. We notice that the normalization constant depends only on C and is, therefore, unaffected by a change in the X representation. Hence, by looking at the trace defining the matrix Bingham density, from the properties of orthogonal matrices it easily follows that the matrix Bingham distribution is invariant with respect to right actions of the orthogonal group:

$$\begin{aligned} p(X'|M,C) &= b(C)^{-1}\text{etr}(CM^\intercal(XQ)(XQ)^\intercal M) && (X'=XQ) \\ &= b(C)^{-1}\text{etr}(CM^\intercal XQQ^\intercal X^\intercal M) && ((XQ)^\intercal=Q^\intercal X^\intercal) \\ &= b(C)^{-1}\text{etr}(CM^\intercal XX^\intercal M) && (QQ^T=I_r, \text{ for } Q\in\mathcal{O}(r)) \\ &= p(X|M,C) && \text{(By definition of matrix Bingham).} \end{aligned} \tag{27}$$

□

B.2 Multimodality

As a result of the former proposition, the matrix Bingham distribution is intrinsically multi-modal. Indeed, it assigns the same density to all the matrices of the form $XQ \in \text{St}(d,r)$, for a given $X \in \text{St}(d,r)$ and for arbitrary $Q \in \mathcal{O}(r)$. The assignment of the same density to all the orthonormal bases related by a right orthogonal transformation is a fundamental property that we leverage to employ the matrix Bingham as a distribution on fixed-dimension subspaces.

B.3 Antipodal symmetry

The invariance of the matrix Bingham density function with respect of right orthogonal transformations implies also antipodal symmetry, i.e., changing in $p(X|M,C)$ the sign of any column of X provides the same density. Indeed, a change of sign in an arbitrary number of columns can be expressed as a right multiplication with a diagonal matrix with only $+1$ or -1 as its diagonal elements, and such a matrix is orthogonal.

Proposition B.2. *The matrix Bingham distribution is antipodal. For all $X \in St(d,r)$ and $S \in \mathbb{R}^{r\times r}$ diagonal matrix with diagonal values in $\{-1,1\}$:*

$$p(X|M,C) = p(XS|M,C). \tag{28}$$

Proof. Let $X \in \text{St}(d,r)$ and $X' = XS$, with $S \in \mathbb{R}^{r\times r}$ an arbitrary diagonal matrix with diagonal entries in $\{-1,1\}$. We notice also in this case that the normalization constant depends only on C and is, therefore, unaffected by a change in the X representation. Since S is a particular orthogonal matrix, the proof follows using the same argument of Proposition B.1:

$$\begin{aligned} p(X'|M,C) &= b(C)^{-1}\text{etr}(CM^\intercal(XS)(XS)^\intercal M) && (X'=XS) \\ &= b(C)^{-1}\text{etr}(CM^\intercal XSS^\intercal X^\intercal M) \\ &= b(C)^{-1}\text{etr}(CM^\intercal XX^\intercal M) && (SS^T=I_r, \text{ for } S\in\mathcal{O}(r)) \\ &= p(X|M,C) && \text{(By definition of matrix Bingham).} \end{aligned} \tag{29}$$

□

B.4 Identifiability

We remark that the Bingham distribution is not identifiable since $p(X|M,C)$ provides the same density when C is replaced with $C' = C + a\mathbb{I}_d$, as the additional constant terms would become part of the normalizing constant. Identifiability can be guaranteed by fixing one of the matrix diagonal values and shifting the other diagonal elements accordingly.

C APPROXIMATION OF THE NORMALIZATION CONSTANT

The density normalizing constant $b(C)$ of a matrix Bingham distribution is a confluent hypergeometric function of matrix argument:

$$\begin{aligned} b(C) &= \int_{X\in\text{St}(d,r)} \text{etr}(CM^\intercal XX^\intercal M)\,\text{d}X \\ &= {}_1F_1\left(\frac{1}{2}r;\frac{1}{2}d;C\right). \end{aligned} \tag{30}$$

Owing to the computational complexity of evaluating a confluent hypergeometric function, the computation of the normalizing constant of the Bingham density can produce a computational bottleneck in Bayesian inferential procedures. Efficient solutions proposed in the literature to approximate ${}_1F_1$ comprise saddle point approximations (Kume et al. (2013)), expansion as series of Jack functions (Koev and Edelman (2006)), Laplace approximation (Butler and Wood (2002)), and look-up tables (Gilitschenski et al. (2015)). Whereas the latter are highly efficient, their use is limited to low-dimensional parameter spaces owing to the intractable number of entries that would be required to achieve sufficient accuracy in high-dimensional parameter spaces.

The Laplace approximation proposed in Butler and Wood (2002), even if highly efficient, limits the relationship between the dimension d and rank r values that can be used for the definition of the Grassmann manifold representing the ambient space. This results from the derivation of the approximation by the integral form of the confluent hypergeometric function of matrix argument. In the case of (10), the conditions result in $r > d-1$ and $r < 1$, preventing its application to the approximation of the matrix Bingham normalizing constant.

In this work, we leverage the approximation of ${}_pF_q$ by its expansion as a truncated series of Jack functions proposed in (Koev and Edelman, 2006) to evaluate the matrix Bingham normalizing constant. Besides approximating the value of the normalizing constant, this formulation allows to compute an approximation of its gradient with respect to the diagonal values of the concentration matrix C. We will show the utility of this in the evaluation of the KL-divergence between two matrix Bingham distributions proposed in Sec. 4.2. For the matrix Bingham normalizing constant in (10), the approximation requires only two scalar parameters besides the matrix argument, which for the Bingham are specified in (10) and lead to:

$$\tilde{b}(C) = {}^{m}_{1}F^{(2)}_{1}\left(\frac{1}{2}r; \frac{1}{2}d; C\right), \tag{31}$$

where the series truncation parameter m provides a trade-off between the approximation accuracy and its computational complexity.

The approximation for the generic hypergeometric function of matrix argument is expressed in (Koev and Edelman, 2006) as:

$$\begin{aligned} &{}^{m}_{p}F^{(\alpha)}_{q}(a_1, \ldots, a_p; b_1, \ldots, b_p; X) = \\ &\sum_{k=0}^{m} \sum_{\kappa \vdash k} \frac{(a_1)^{(\alpha)}_{\kappa} \cdots (a_p)^{(\alpha)}_{\kappa}}{k!(b_1)^{(a)}_{\kappa} \cdots (b_q)^{(\alpha)}_{\kappa}} \cdot C^{(\alpha)}_{\kappa}(X), \end{aligned} \tag{32}$$

where $p, q \geq 0$ are integers, $\alpha > 0$ is a parameter, and $X \in \mathbb{R}^{n \times n}$ is symmetric matrix and the truncation is computed for all partitions $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_\ell)$ of k s.t. $|\kappa| \leq m$. $|\kappa| \coloneqq \kappa_1 + \kappa_2 + \cdots + \kappa_\ell$, $C^{(\alpha)}_{\kappa}(X)$ is the Jack function, and $(a)^{(\alpha)}_{\kappa}$ is the generalized Pochhammer symbol:

$$(a)^{(\alpha)}_{\kappa} = \prod_{(i,j) \in \kappa} \left(a - \frac{i-1}{\alpha} + j - 1\right).$$

The series truncation parameter m provides a trade-off between the approximation accuracy and its computational complexity. ${}_pF^{(\alpha)}_q$ depends on X is only through its eigenvalues. Therefore, it can be considered as diagonal without loss of generality. We refer the reader to (Koev and Edelman, 2006) for a detailed description of the approximation algorithm and its computational complexity.

For the matrix Bingham normalizing constant in (30), we consider the case for $p = 1$, $q = 1$ and $\alpha = 2$ of (32). In this case, the approximation requires only two scalar parameters besides the matrix argument, which for the Bingham are specified in (30) and lead to the following approximation:

$$\tilde{b}(C) = {}^{m}_{1}F^{(2)}_{1}\left(\frac{1}{2}r; \frac{1}{2}d; C\right). \tag{33}$$

C.1 Discussion on the truncation error

We have extended our experiments focusing on the analysis of the truncation error of the matrix Bingham normalizing constant approximation, which can affect the computation of the geodesics through the optimization of the KL divergence. As expected, the approximation error significantly decreases with the increase of the

truncation order m, where we considered the value of the hypergeometric function of matrix argument for $m = 20$ as reference comparison term since it has led to convergence in all the examined cases. Since, in general, the approximation error depends on the distribution of the input concentration values, we examined the impact of truncation order selection on the geodesics computation task, finding that a truncation order from $m = 3$ to $m = 6$ provided meaningful geodesics crossing the latent data domain in all our experiments. We considered this range of truncation order as feasible with the current implementation, which is based on JAX automatic differentiation to determine the gradient of the matrix Bingham normalizing constant required to compute the KL-divergence. In order to speed up computations, we used the just-in-time (JIT) compilation feature of JAX, caching the compilation to load it directly from disk over consecutive runs. This has highly improved the computational performance during model training and the evaluation of KL-divergence and Fisher-Rao metric.

D COMPUTATION OF THE FISHER INFORMATION MATRIX OF A MATRIX BINGHAM RANDOM VARIABLE

In this section, we detail the computation of the Fisher information matrix for a matrix Bingham distribution with probability density function $p(X|M, C)$, where $X \in \mathrm{St}(d, r)$, $M \in \mathrm{St}(d, k)$ is a d-dimensional k-frame, $C \in \mathbb{R}^{k\times k}$ is a diagonal concentration matrix, and $1 \leq k \leq d$ is a hyperparameter of the distribution accounting for the low-rank modeling of M. When $k = d$, $M \in \mathcal{O}(d)$ reduces to an orthogonal matrix and C becomes a $d \times d$ diagonal matrix. In the following, we will consider the Stiefel manifold as a Riemannian submanifold of the embedding space $\mathbb{R}^{d\times k}$.

Let $\mathbf{m} = \mathrm{vec}(M)$ and $\mathbf{c} = \mathrm{diag}(C)$. The Fisher information matrix $\mathbf{I}_{\mathcal{H}}$ of a random variable $X \sim p(X|M, C)$ is given by minus the expected value of the Hessian of the log-likelihood with respect to the distribution parameters:

$$[\mathbf{I}_{\mathcal{H}}(\theta)]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\theta_j}\log p(X|\theta)\right], \tag{34}$$

where $\theta = [\mathbf{m}, \mathbf{c}]$ denotes the vector of the distribution parameters. This holds under the required differentiability assumptions for the log-likelihood function w.r.t. its parameters, which are satisfied in the considered case.

The differentiation problem can be decomposed into 5 subproblems, developed in the following subsections:

1. Differentiation of $\log p(X|M, C)$ w.r.t. M;
2. Differentiation of $\log p(X|M, C)$ w.r.t. C;
3. Differentiation of $\log p(X|M, C)$ twice w.r.t. M;
4. Differentiation of $\log p(X|M, C)$ twice w.r.t. $\mathrm{diag}(C)$;
5. Differentiation of $\log p(X|M, C)$ w.r.t. M and w.r.t. $\mathrm{diag}(C)$;

Under the decomposition of the distribution parameters in θ, the resulting Fisher information matrix can be written as the following block matrix:

$$\mathbf{I}_{\mathcal{H}}(\theta) = \begin{bmatrix} \mathbf{I}_{\mathcal{H},\mathbf{mm}} & \mathbf{I}_{\mathcal{H},\mathbf{mc}} \\ \mathbf{I}_{\mathcal{H},\mathbf{cm}} & \mathbf{I}_{\mathcal{H},\mathbf{cc}} \end{bmatrix}, \tag{35}$$

where $\mathbf{I}_{\mathcal{H},\mathbf{xy}}$ indicates second-order differentiation of the negative log-likelihood w.r.t. the corresponding sets of parameters (among $\mathbf{m}$ and $\mathbf{c}$). Since the Fisher information matrix $\mathbf{I}_{\mathcal{H}}(\theta)$ is symmetric, the derivation of only one of the two non-diagonal blocks in (35) is required, as the other can be obtained as its transpose, i.e., $\mathbf{I}_{\mathcal{H},\mathbf{cm}} = (\mathbf{I}_{\mathcal{H},\mathbf{mc}})^\intercal$.

We notice that M is naturally defined over the Stiefel manifold $\mathrm{St}(d, k)$. Therefore, in the following, when differentiating w.r.t. M, we consider the Riemannian gradient and the Riemannian Hessian as natural gradient and Hessian over the Stiefel manifold $\mathrm{St}(d, k)$, respectively. Moreover, we consider the Stiefel manifold $\mathrm{St}(d, k)$ as a submanifold of $\mathbb{R}^{d\times k} \cong \mathbb{R}^{d\cdot k}$, inheriting its standard inner product.

Referring to (Boumal, 2023), we report the propositions for Riemannian gradient and Riemannian Hessian of a real function defined on a Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathcal{E}$.

Proposition D.1. *Let $\mathcal{M}$ be a Riemannian submanifold of $\mathcal{E}$ endowed with the metric $\langle\cdot,\cdot\rangle$ and let $f:\mathcal{M}\to\mathbb{R}$ be a smooth function. The Riemannian gradient of f is given by*

$$grad f(x) = Proj_x(grad \bar{f}(x)), \tag{36}$$

where $\bar{f}$ is any smooth extension of f to a neighborhood of $\mathcal{M}$ in $\mathcal{E}$, and $Proj_x(\cdot)$ is the orthogonal projection onto the tangent space of $\mathcal{M}$ at x.

For a Stiefel manifold endowed with the metric derived from its embedding Euclidean space, the Riemannian gradient can be expressed as

$$\mathrm{grad} f(X) = \mathrm{Proj}_X(\mathrm{grad}\bar{f}) = \mathrm{grad}\bar{f}(X) - \frac{1}{2}X\left[X^\intercal \mathrm{grad}\bar{f}(X) + \mathrm{grad}\bar{f}(X)^\intercal X\right]. \tag{37}$$

where $\mathrm{Proj}_X : \mathrm{St}(d,k) \to \mathrm{T}_X\mathrm{St}(d,k)$ is specified by

$$\mathrm{Proj}_X(U) = U - X\frac{X^TU + U^TX}{2}. \tag{38}$$

Proposition D.2. *Let $\mathcal{M}$ be a Riemannian submanifold of $\mathcal{E}$ endowed with the metric $\langle\cdot,\cdot\rangle$ and let $f:\mathcal{M}\to\mathbb{R}$ be a smooth function. Let $\bar{G}$ be a smooth extension of grad$f$. Then,*

$$Hess f(x)[u] = Proj_x(\mathrm{D}\bar{G}(x)[u]). \tag{39}$$

In the case of a Stiefel manifold, the smooth extension of grad$f$ is given by

$$\bar{G}(X) = \mathrm{grad}\bar{f}(X) - \frac{1}{2}X\left[X^\intercal \mathrm{grad}\bar{f}(X) + \mathrm{grad}\bar{f}(X)^\intercal X\right], \tag{40}$$

where the domain of $\bar{G}$ extends in a neighborhood of $\mathrm{St}(d,k)$ in $\mathbb{R}^{d\times k}$.

The Riemannian Hessian of $f:\mathrm{St}(d,k)\to\mathbb{R}$ in X along V is

$$\mathrm{Hess} f(X)[V] = \mathrm{Proj}_x\left(\mathrm{Hess}\bar{f}(X)[V] - \frac{1}{2}V\left[X^\intercal \mathrm{grad}\bar{f}(X) + \mathrm{grad}\bar{f}(X)^\intercal X\right]\right). \tag{41}$$

D.1 First-order differentiation of $\log p(X|M,C)$ w.r.t. M

We start by performing the differentiation with respect to M of a smooth extension $\bar{f}(M)$ of $f(M) = \log p(X|M,C)$ in $\mathcal{E}$, rewriting the log-likelihood in a more suitable form and using the rules of matrix differentiation. Let $B = XX^\intercal$.

$$\begin{aligned}
\frac{\partial}{\partial M}\log p(X|M,C) &= \frac{\partial}{\partial M}\left(-\log b(C) + \mathrm{tr}(CM^TXX^TM)\right)\\
&= \frac{\partial}{\partial M}(-\log b(C)) + \frac{\partial}{\partial M}\mathrm{tr}(CM^TXX^TM)\\
&= \frac{\partial}{\partial M}\mathrm{tr}(CM^TXX^TM)\\
&= \frac{\partial}{\partial M}\mathrm{tr}(MCM^TXX^T)\\
&= \frac{\partial}{\partial M}\mathrm{tr}(MCM^TB) \qquad\qquad (B=XX^\intercal)\\
&= \frac{\partial}{\partial M}\mathrm{tr}(M^TBMC)\\
&= BMC + B^TMC^T\\
&= XX^TMC + XX^TMC^T = 2XX^TMC.
\end{aligned} \tag{42}$$

Then, we compute the Riemannian gradient of $f(M) = \log p(X|M,C)$ w.r.t. M by specializing (37):

$$G = \mathrm{grad}\bar{f}(M) = 2XX^TMC, \tag{43}$$

$$\mathrm{grad} f(M) = G - \frac{1}{2}M(M^TG + G^TM). \tag{44}$$

D.2 First-order differentiation of $\log p(X|M,C)$ w.r.t. C

We differentiate $\log p(X|M,C)$ w.r.t. C as:

$$\begin{aligned}\frac{\partial}{\partial C}\log p(X|M,C) &= \frac{\partial}{\partial C}\left(-\log b(C) + \operatorname{tr}(CM^TXX^TM)\right)\\ &= -\frac{\partial}{\partial C}(\log b(C)) + \frac{\partial}{\partial C}\operatorname{tr}(CM^TXX^TM)\\ &= -\frac{1}{b(C)}\frac{\partial(b(C))}{\partial C} + (M^TXX^TM)\circ\mathbb{I}_k,\end{aligned} \tag{45}$$

where $A\circ B$ denotes the Hadamard product between matrices A and B, and $\mathbb{I}_k$ is the identity matrix of order k.

The derivative of $\log p(X|M,C)$ with respect to $\operatorname{diag}(C)$ is the diagonal of the derivative w.r.t. C:

$$\frac{\partial}{\partial \operatorname{diag}(C)}\log p(X|M,C) = \operatorname{diag}\left(\frac{\partial}{\partial C}\log p(X|M,C)\right). \tag{46}$$

D.3 Second-order differentiation of $\log p(X|M,C)$ w.r.t. M

To determine the Riemannian Hessian of $f(M) = \log p(X|M,C)$ w.r.t. $M \in \mathrm{St}(d,k)$ along $V \in \mathrm{T}_M\mathrm{St}(d,k)$, we specialize (41):

$$\operatorname{Hess} f(M)[V] = \operatorname{Proj}_M\left(\tilde{H} : V - \frac{1}{2}V(M^\intercal G + G^\intercal M)\right) \tag{47}$$

where $A:B$ is the diadic product between the 4^{th}-order tensor A and the 2^{nd}-order tensor B, $\tilde{H} = \operatorname{Hess}\bar{f}(M)[V]$ is the Hessian of the smooth extension of f in $\mathcal{E}$, G is specified in (43), and Proj_M is the orthogonal projection to the tangent space of the Stiefel manifold at M as defined in (38).

By splitting the 4 indices with respect to the two indices of the first differentiation parameter and the two indices of the second one, the elements of the 2nd-order tensor resulting from the diadic product $D = \tilde{H} : V$ can be expressed as

$$(D)_{ij} = \sum_{m\ell}\tilde{H}_{(ij)(m\ell)}V_{m\ell}. \tag{48}$$

The Hessian $\operatorname{Hess}\bar{f}(M)[V]$ of the smooth extension of f in $\mathcal{E}$ w.r.t. M is a 4th-order tensor whose elements are given by

$$\begin{aligned}\tilde{H}_{(ij)(m\ell)} &= 2\frac{\partial}{\partial M_{m\ell}}\sum_{k,s}P_{ik}M_{ks}C_{sj}\\ &= 2\sum_{k,s}P_{ik}\delta_{mk}\delta_{\ell s}C_{sj}\\ &= 2\sum_{k}P_{ik}\delta_{mk}C_{\ell j}\\ &= 2P_{im}C_{\ell j}\\ &= 2P_{im}\delta_{\ell j}\mathbf{c}_j,\end{aligned} \tag{49}$$

where the replacement $P = XX^\intercal$ has been used to simplify the expression, and δ_{ij} is the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$, else $\delta_{ij} = 0$).

To materialize the Riemannian Hessian matrix, required for the computation of the Fisher information matrix $\mathbf{I}_{\mathcal{H}}$, we replaced each of the elements of the canonical basis of $\mathbb{R}^{d\times k}$ as V in (47).

$$H^M_{ij} = \operatorname{Hess} f(M)[E_{ij}], \tag{50}$$

where $E_{ij} \in \mathbb{R}^{d\times k}$ is a matrix unit, i.e., a single-entry matrix with a 1 in the entry indexed by the ith row and the jth column, for $i \in \{1,\cdots,d\}$ and $j \in \{1,\cdots,k\}$. Then, we reshaped the H tensor to obtain the 2^{nd}-order representation $\bar{H}$ used to construct $\mathbf{I}_{\mathcal{H}}$:

$$\bar{H}^{\mathbf{m}}_{(j-1)d+i,(\ell-1)d+m} = H^M_{(ij)(m\ell)}, \tag{51}$$

for $i, m \in \{1, \dots, d\}$ and $j, \ell \in \{1, \dots, k\}$.

D.4 Second-order differentiation of $\log p(X|M, C)$ w.r.t. C

We derive here the second-order differentiation of $\log p(X|M, C)$ w.r.t. C as the following 4th-order tensor:

$$\begin{aligned} H^C = \frac{\partial^2}{\partial C^2} \log p(X|M, C) &= \frac{\partial}{\partial C} \left(-\frac{1}{b(C)} \frac{\partial (b(C))}{\partial C} + (M^T X X^T M) \circ \mathbb{I}_k \right) \\ &= -\left(-\frac{\frac{\partial}{\partial C} b(C)}{b(C)^2} \frac{\partial}{\partial C} b(C) + \frac{1}{b(C)} \frac{\partial^2 b(C)}{\partial C^2} \right) \\ &= \frac{\frac{\partial}{\partial C} b(C)}{b(C)^2} \frac{\partial}{\partial C} b(C) - \frac{1}{b(C)} \frac{\partial^2 b(C)}{\partial C^2}. \end{aligned} \tag{52}$$

We notice that this 4th-order tensor reduces to a matrix when only the indices related to the diagonal of C are considered:

$$\bar{H}^{\mathbf{c}}_{ij} = H^C_{(ii)(jj)}. \tag{53}$$

This derivation highlights that both the first-order and the second-order derivatives of the matrix Bingham normalization constant $b(C)$ are required to differentiate $\log p(X|M, C)$ twice w.r.t. C. In Appendix H, we discuss the numerical differentiation of the normalization constant $b(C)$ by means of auto-differentiation tools.

D.5 Differentiation of $\log p(X|M, C)$ w.r.t. M and diag(C)

To differentiate w.r.t. diag(C) the Riemannian gradient of $\log p(X|M, C)$, we first expand (44) as:

$$\begin{aligned} \text{grad} f(M) &= G - \frac{1}{2} M (M^T G + G^T M) \\ &= 2XX^T MC - \frac{1}{2} M (M^T 2XX^T MC + (2XX^T MC)^T M) \\ &= 2XX^T MC - M(M^T XX^T MC + CM^T XX^T M) \\ &= 2XX^T MC - MM^T XX^T MC - MCM^T XX^T M \end{aligned} \tag{54}$$

where G has been replaced with $2XX^T MC$ as derived in (43).

Let $\mathbf{c} = \text{diag}(C)$, $\ell \in 1, \dots, k$, and $P = XX^\intercal$. We differentiate the three obtained terms w.r.t. $\mathbf{c}$ as:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{c}_\ell} (MM^T XX^T MC)_{ij} &= \frac{\partial}{\partial \mathbf{c}_\ell} \sum_{m,n,s,r} M_{im} M_{nm} P_{ns} M_{sr} C_{rj} \\ &= \frac{\partial}{\partial \mathbf{c}_\ell} \sum_{m,n,s,r} M_{im} M_{nm} P_{ns} M_{sr} \delta_{rj} \mathbf{c}_j \\ &= \sum_{m,n,s,r} M_{im} M_{nm} P_{ns} M_{sr} \delta_{rj} \delta_{\ell j} \\ &= \sum_{m,n,s} M_{im} M_{nm} P_{ns} M_{sj} \delta_{\ell j} \\ &= \delta_{\ell j} \sum_{m,n,s} M_{im} M_{nm} P_{ns} M_{sj} \\ &= \delta_{\ell j} [MM^T PM]_{ij}, \end{aligned} \tag{55}$$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{c}_\ell}(2XX^TMC)_{ij} &= \frac{\partial}{\partial \mathbf{c}_\ell}\sum_{m,n} 2P_{im}M_{mn}C_{nj} \\
&= \frac{\partial}{\partial \mathbf{c}_\ell}\sum_{m,n} 2P_{im}M_{mn}\delta_{nj}\mathbf{c}_j \\
&= \frac{\partial}{\partial \mathbf{c}_\ell}\sum_{m} 2P_{im}M_{mj}\mathbf{c}_j \\
&= \delta_{j\ell}\sum_{m} 2P_{im}M_{mj} \\
&= 2\delta_{j\ell}[PM]_{ij},
\end{aligned} \tag{56}$$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{c}_\ell}(MCM^TXX^TM)_{ij} &= \frac{\partial}{\partial \mathbf{c}_\ell}\sum_{m,n,s,t} M_{im}C_{mn}M_{sn}P_{st}M_{tj} \\
&= \frac{\partial}{\partial \mathbf{c}_\ell}\sum_{m,n,s,t} M_{im}\delta_{mn}\mathbf{c}_n M_{sn}P_{st}M_{tj} \\
&= \sum_{m,n,s,t} M_{im}\delta_{mn}\delta_{n\ell}M_{sn}P_{st}M_{tj} \\
&= \sum_{n,s,t} M_{in}\delta_{n\ell}M_{sn}P_{st}M_{tj} \\
&= \sum_{s,t} M_{i\ell}M_{s\ell}P_{st}M_{tj} \\
&= M_{i\ell}\sum_{s,t}(M^T)_{\ell s}P_{st}M_{tj} \\
&= M_{i\ell}[M^TPM]_{\ell j}.
\end{aligned} \tag{57}$$

The resulting Hessian matrix is defined by

$$H^{M\mathbf{c}}_{(ij)\ell} = \left(\frac{\partial \mathrm{grad} f(M)}{\partial \mathbf{c}}\right)_{(ij)\ell} = \delta_{\ell j}[MM^TPM]_{ij} + 2\delta_{j\ell}[PM]_{ij} + M_{i\ell}[M^TPM]_{\ell j}. \tag{58}$$

After compressing the dimensions related to the differentiation w.r.t. M into a single dimension, the obtained 3rd-order tensor results in the following matrix:

$$\bar{H}^{\mathbf{mc}}_{(j-1)d+i,\ell} = H^{M\mathbf{c}}_{(ij)\ell}. \tag{59}$$

D.6 Expected value of the Hessian matrices w.r.t. the distribution parameters

Since all the derived Hessian matrices depend on X only through the form $P = XX^\intercal$, and given that they depend on P linearly, the computation of the expected value of the Hessian matrices reduces to the computation of $\mathbb{E}[XX^\intercal]$. This derivation is developed in Appendix F and is shown in (64) to result into the following expression:

$$\mathbb{E}[XX^T] = M\frac{1}{b(C)}\frac{\partial b(C)}{\partial C}M^T. \tag{60}$$

The matrices $\mathbf{I}_{\mathcal{H},\mathbf{mm}}$, $\mathbf{I}_{\mathcal{H},\mathbf{cc}}$ and $\mathbf{I}_{\mathcal{H},\mathbf{mc}}$ can then be computed through the negated expected value of the obtained and reshaped Hessian matrices:

$$\begin{aligned}
\mathbf{I}_{\mathcal{H},\mathbf{mm}} &= -\mathbb{E}[\bar{H}^{\mathbf{m}}], \\
\mathbf{I}_{\mathcal{H},\mathbf{cc}} &= -\mathbb{E}[\bar{H}^{\mathbf{c}}], \\
\mathbf{I}_{\mathcal{H},\mathbf{mc}} &= -\mathbb{E}[\bar{H}^{\mathbf{mc}}].
\end{aligned} \tag{61}$$

E KL-DIVERGENCE BETWEEN MATRIX BINGHAM DISTRIBUTIONS

In this section, we detail the derivation of the KL-divergence between two matrix Bingham distributions.

Proposition E.1. *The KL-divergence between two matrix Bingham distributions P and Q with probability density functions $p(X|M_1, C_1)$ and $q(X|M_2, C_2)$, respectively, is*

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= \int_{St(d,r)} p(X|M_1, C_1) \log\left(\frac{p(X|M_1, C_1)}{q(X|M_2, C_2)}\right) \mathrm{d}X \\ &= \log\frac{b(C_2)}{b(C_1)} + Tr((M_1 C_1 M_1^\intercal - M_2 C_2 M_2^\intercal)\mathbb{E}_p[XX^\intercal]), \end{aligned} \tag{62}$$

where the integral is performed over the Stiefel manifold $St(d, r)$ and the normalizing constants $b(C)$ and $b(V)$ are defined as in (10).

Proof. Let $G = M_1 C_1 M_1^T$ and $H = M_2 C_2 M_2^T$. Then:

$$\begin{aligned} D_{\mathrm{KL}} &= \int_{St(d,r)} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \\ &= \int_{St(d,r)} p(X|G) \log\left(\frac{p(X|G)}{q(X|H)}\right) \mathrm{d}X \\ &= \int_{St(d,r)} \frac{1}{b(G)}\mathrm{etr}(X^T G X) \log\left(\frac{\frac{1}{b(G)}\mathrm{etr}(X^T G X)}{\frac{1}{b(H)}\mathrm{etr}(X^T H X)}\right) \mathrm{d}X \\ &= \int_{St(d,r)} \frac{1}{b(G)}\mathrm{etr}(X^T G X) \left(\log\frac{1}{b(G)}\mathrm{etr}(X^T G X) - \log\frac{1}{b(H)}\mathrm{etr}(X^T H X)\right) \mathrm{d}X \\ &= \int_{St(d,r)} \frac{1}{b(G)}\mathrm{etr}(X^T G X) \left(-\log b(G) + tr(X^T G X) + \log b(H) - tr(X^T H X)\right) \mathrm{d}X \\ &= \int_{St(d,r)} \frac{1}{b(G)}\mathrm{etr}(X^T G X) \left(\log\frac{b(H)}{b(G)} + tr(X^T G X) - tr(X^T H X)\right) \mathrm{d}X \\ &= \log\frac{b(H)}{b(G)} + \int_{St(d,r)} \frac{1}{b(G)}\mathrm{etr}(X^T G X) \left(tr(X^T G X) - tr(X^T H X)\right) \mathrm{d}X \\ &= \log\frac{b(H)}{b(G)} + \mathbb{E}_1[tr(X^T G X) - tr(X^T H X)] \\ &= \log\frac{b(H)}{b(G)} + \mathbb{E}_1[tr(G X X^T) - tr(H X X^T)] \\ &= \log\frac{b(H)}{b(G)} + \mathbb{E}_1[tr(G X X^T - H X X^T)] \\ &= \log\frac{b(H)}{b(G)} + \mathbb{E}_1[tr((G - H) X X^T)] \\ &= \log\frac{b(H)}{b(G)} + tr((G - H)\mathbb{E}_1[X X^T]) \\ &= \log\frac{b(H)}{b(G)} + tr\left((G - H) M_1 \frac{1}{b(C_1)} \frac{\partial b(C_1)}{\partial C_1} M_1^T\right) \\ &= \log\frac{b(C_2)}{b(C_1)} + tr\left(\left(M_1 C_1 M_1^T - M_2 C_2 M_2^T\right) M_1 \frac{1}{b(C_1)} \frac{\partial b(C_1)}{\partial C_1} M_1^T\right) \end{aligned} \tag{63}$$

□

The derived KL-divergence depends on the difference of the exponential trace arguments of the two matrix Bingham distributions, on the expected value of the correlation matrix, and on the ratio of the normalization constants.

F COMPUTATION OF THE EXPECTED VALUE OF $XX^\intercal$ WITH RESPECT TO A MATRIX BINGHAM DISTRIBUTION

In this section, we derive the expected value of the random variable $XX^\intercal$ for a matrix Bingham distribution with probability density function $p(X|M,C)$, where $X \in \mathrm{St}(d,r)$, $M \in \mathrm{St}(d,k)$ is a d-dimensional k-frame, $C \in \mathbb{R}^{k\times k}$ is a diagonal concentration matrix:

$$
\begin{aligned}
\mathbb{E}[XX^T] &= \int_{St(d,r)} XX^T \frac{1}{b(C)} \mathrm{etr}(CM^TXX^TM)dX \\
&= \frac{1}{b(C)} \int_{St(d,r)} XX^T \mathrm{etr}(CM^TXX^TM)dX \\
&= \frac{1}{b(C)} M \int_{St(d,r)} M^TXX^T \mathrm{etr}(CM^TXX^TM)MdXM^T \\
&= \frac{1}{b(C)} M \int_{St(d,r)} M^TXX^TM \,\mathrm{etr}(CM^TXX^TM)dXM^T \\
&= \frac{1}{b(C)} M \int_{St(d,r)} \frac{\partial}{\partial C} \mathrm{etr}(CM^TXX^TM)dXM^T \\
&= \frac{1}{b(C)} M \frac{\partial}{\partial C} \int_{St(d,r)} \mathrm{etr}(CM^TXX^TM)dXM^T \\
&= \frac{1}{b(C)} M \frac{\partial}{\partial C} b(C) M^T \\
&= M \frac{1}{b(C)} \frac{\partial b(C)}{\partial C} M^T \\
&= M \,\mathrm{diag}\left(\frac{1}{b(C)} \frac{\partial b(C)}{\partial \mathbf{c}_1}, \cdots, \frac{1}{b(C)} \frac{\partial b(C)}{\partial \mathbf{c}_d} \right) M^T
\end{aligned}
\tag{64}
$$

We notice that, under the derivation assumptions, this expression holds for $M \in \mathcal{O}(d)$. When $M \in \mathrm{St}(d,k)$, with $1 \leq k \leq d, k \in \mathbb{N}$, the completion of M with its orthogonal complement in $\mathbb{R}^{d\times d}$ is required to have $M \in \mathcal{O}(d)$.

G HYPERPARAMETRIZATION OF THE SUBSPACE DIMENSIONS

The direction matrix M and the diagonal concentration matrix C in the matrix Bingham likelihood (9) and in the related equations discussed above can be consistently reduced such that $M \in \mathrm{St}(d,r)$ has rank $r \leq d$ and $C \in \mathbb{R}^{r\times r}$ is a diagonal matrix reduced accordingly. This amounts to possibly setting $d-r$ elements of the diagonal of C to 0 in the original formulations. Besides lowering the computational complexity, this provides a further hyperparameter to flexibly manage the output probabilistic modeling based on prior knowledge on the subspaces featuring the data.

This can be particularly relevant when the data is known to be low-rank (i.e., $r \ll d$) in the ambient space, and a reduced number of considered directions in the matrix Bingham can be useful to model the output distribution accordingly. Therefore, in the proposed experiments, we will consider the number of columns of M and diagonal elements of C as an hyperparameter to be determined by means of hyperparameter search or to be set based on prior knowledge on the dimension of the subspaces featuring the input data.

H DIFFERENTIATING THE NORMALIZATION CONSTANT

We implemented the computation of the matrix Bingham normalization constant approximation based on (Koev and Edelman, 2006) in JAX to allow its automatic differentiation with respect to the concentration parameters. We then used the Jax2Torch[2] library to enable the end-to-end optimization of the implemented PyTorch model. We used this procedure both for training purposes and for the computation of the gradient and Hessian of the approximated normalization constant, required to compute the Fisher information matrix for the matrix Bingham and the pull-back metric.

[2] https://github.com/lucidrains/jax2torch

I MODE OF A MATRIX BINGHAM DISTRIBUTION

We show in the following that the mode of a matrix Bingham distribution can be easily derived as the leading eigenvectors of the directional parameter matrix. This is an extension of the result obtained for the Bingham distribution defined on the n-sphere, where the mode is also provided by the leading eigenvector of the directional parameter matrix.

Proposition I.1. *Let $p_{\mathcal{B}}(X|M,C)$ be a matrix Bingham distribution as in (9), with $X \in St(d,r)$, $M \in O(d)$ and $C = diag(\mathbf{c}), \mathbf{c} \in \mathbb{R}^d_+$. A mode of the matrix Bingham distribution is given by the leading eigenvectors of matrix $MCM^\intercal$.*

Proof. By definition, the matrix Bingham pdf is given, as in (9), by

$$p_{\mathcal{B}}(X|M,C) = b(C)^{-1}\text{etr}(CM^\intercal XX^\intercal M). \tag{65}$$

We rewrite the latter as

$$p(X|G) = b(G)^{-1}\text{etr}(GXX^\intercal), \tag{66}$$

with $G = MCM^\intercal$ symmetric positive semidefinite.

To determine the mode of the distribution, we aim at determining the $X \in \text{St}(d,r)$ that maximizes (66) given M and C. We equivalently aim to maximize the logarithm of (66) to simplify derivations:

$$\arg\max_{X\in\text{St}(d,r)} -\log b(G) + \text{Tr}(GXX^\intercal). \tag{67}$$

Since $\log b(G)$ is constant, we can write the optimization problem as:

$$\arg\max_{X\in\text{St}(d,r)} \text{Tr}(GXX^\intercal). \tag{68}$$

Let $P = XX^\intercal$. Both P and G are symmetric positive semidefinite. Therefore, from a corollary of Von Neumann's trace inequality, we obtain:

$$\text{Tr}(GXX^T) = \text{Tr}(GP) \leq \sum_{i=1}^{d} g_i p_i, \tag{69}$$

where p_i and g_i, $i = 1,\dots,d$ are the eigenvalues of P and G, respectively, sorted in decreasing order.

Since P is a projection matrix, its eigenvalues are such that $p_i \in \{0,1\}$ for $i = 1,\dots,d$ and $\sum_{i=0}^{d} p_i = r$. Moreover, given the decreasing sorting of the eigenvalues, the p_i's are such that $p_i = 1$ for $i = 1,\dots,r$ and $p_i = 0$ for $i = r+1,\dots,d$. Therefore, the upper bound derived from the Von Neumann's trace inequality can be reduced to

$$\text{Tr}(GP) \leq \sum_{i=1}^{d} g_i p_i = \sum_{i=1}^{r} g_i p_i, \tag{70}$$

where the sum on the RHS is now up to r instead of d.

Let M_r be the r leading eigenvectors of G, given by the columns of M corresponding to the r highest entries of C's diagonal. Let $X = M_r$. By expanding G and P, and using the cyclic permutation property of the trace, we obtain

$$\text{Tr}(GP) = \text{Tr}(MCM^\intercal M_r M_r^\intercal) = \text{Tr}(CM^\intercal M_r M_r^\intercal M). \tag{71}$$

We assume without loss of generality that $M[:,:r] = M_r$, i.e., M_r coincides with the leading columns of M. This can be easily obtained by permuting the columns of M and the corresponding entries in C's diagonal so that the first r columns of M are associated to the highest entries of C. With this assumption, it results that

$$M^\intercal M_r M_r^\intercal M = M^\intercal M_r (M^\intercal M_r)^\intercal \tag{72}$$

$$= \begin{bmatrix} \mathbb{I}_r \\ \mathbf{0}_{d-r\times r} \end{bmatrix} \begin{bmatrix} \mathbb{I}_r & \mathbf{0}_{r\times(d-r)} \end{bmatrix} \tag{73}$$

$$= \begin{bmatrix} \mathbb{I}_r & \mathbf{0}_{r\times(d-r)} \\ \mathbf{0}_{(d-r)\times r} & \mathbf{0}_{(d-r)} \end{bmatrix} \tag{74}$$

Then,

$$\begin{aligned}\mathrm{Tr}(GP) &= \mathrm{Tr}(CM^\intercal M_r M_r^\intercal M) && (75)\\ &= \mathrm{Tr}\left(C\begin{bmatrix}\mathbb{I}_r & \mathbf{0}_{r\times(d-r)}\\ \mathbf{0}_{(d-r)\times r} & \mathbf{0}_{(d-r)}\end{bmatrix}\right) && (76)\\ &= \mathrm{Tr}(C_r) = \sum_{i=1}^{r} g_i, && (77)\end{aligned}$$

where C_r is the submatrix of C given by its first r rows and columns.

Therefore, the choice $X = M_r$ maximizes (68) as it fulfills the equality in (70). □

This proposition holds also under the following conditions:

- When G is low-rank, i.e., $M \in \mathrm{St}(d, r')$, $C = \mathrm{diag}(\mathbf{c}), \mathbf{c} \in \mathbb{R}_+^{r'}$, and $r' \geq r$, since the proof leverages only the leading eigenvectors of M and does not require that G be full-rank.
- In the case of complex-valued matrix Bingham distributions, since (69) also holds for Hermitian positive semidefinite matrices, providing an analogous upper bound as for the former proof in the real-valued case.

J REAL-VALUED REPRESENTATION OF A COMPLEX-VALUED MATRIX BINGHAM DISTRIBUTION

In this Appendix, we provide the representation of a complex-valued matrix Bingham distribution as a real-valued one with suitably structured parameter matrices. First, we derive a real-valued matrix representation for the complex-valued distribution parameters. Then, we show that using real-valued parameter matrices with a given structure produces a real-valued matrix Bingham distribution that is equivalent to the complex-valued matrix Bingham.

Let $\mathrm{St}_{\mathbb{C}}(d, r) = \{X \in \mathbb{C}^{d\times r} \,|\, X^H X = \mathbb{I}_r\}$ be the complex Stiefel manifold, and $\mathrm{U}(d) = \{U \in \mathbb{C}^{d\times d} \,|\, U^H U = UU^H = \mathbb{I}_d\}$ be the unitary group. A complex-valued matrix Bingham distribution is defined analogously to the real-valued one as:

$$p_{\mathcal{CB}}(X|\tilde{M}, \tilde{C}) = b(\tilde{C})^{-1}\mathrm{etr}(\tilde{C}\tilde{M}^H X X^H \tilde{M}), \tag{78}$$

where $\mathrm{etr}(\,\cdot\,)$ is the exponential of the trace of a square matrix, $X \in \mathrm{St}_{\mathbb{C}}(d, r)$, $\tilde{M} \in \mathrm{U}(d)$, and $\tilde{C} \in \mathbb{R}^d$ is diagonal. $b(\tilde{C})$ is the normalizing constant of the distribution and, as for the real case, it depends only on the $\tilde{C}$ concentrations matrix.

As indicated by Kent (1994) and Bingham et al. (1992) for the hypersphere case, the complex matrix Bingham distribution can be represented by a real-valued Bingham distribution over a twice dimensional space.

Proposition J.1. *Let $\tilde{X} \in St_{\mathbb{C}}(d, r)$ and $X \in St(2d, 2r)$. The complex-valued matrix Bingham distribution $p_{\mathcal{CB}}(\tilde{X}|\tilde{M}, \tilde{C})$ is equivalent to a real-valued matrix Bingham distribution $p_{\mathcal{B}}(X|M, C)$ s.t.*

$$X = \begin{bmatrix}\tilde{X}_r & -\tilde{X}_i\\ \tilde{X}_i & \tilde{X}_r\end{bmatrix} \in \mathbb{R}^{2d\times 2r}, \quad M = \begin{bmatrix}\tilde{M}_r & -\tilde{M}_i\\ \tilde{M}_i & \tilde{M}_r\end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \quad C = \begin{bmatrix}\tilde{C} & \mathbf{0}\\ \mathbf{0} & \tilde{C}\end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \tag{79}$$

where $A_r = \mathrm{Re}(A)$ and $A_i = \mathrm{Im}(A)$ denote, respectively, the real and imaginary parts of a generic matrix A.

Proof. Let $\tilde{G} = \tilde{M}\tilde{C}\tilde{M}^H \in \mathbb{C}^{d\times d}$, and let $\tilde{G}_r = \mathrm{Re}(\tilde{G})$ and $\tilde{G}_i = \mathrm{Im}(\tilde{G})$ be the real and imaginary parts of $\tilde{G}$, respectively. Under multiplication between matrices, the complex-valued matrix $\tilde{G}$ and the random matrix $\tilde{X}$ can be equivalently represented by the real-valued matrices:

$$X = \begin{bmatrix}\tilde{X}_r & -\tilde{X}_i\\ \tilde{X}_i & \tilde{X}_r\end{bmatrix} \in \mathbb{R}^{2d\times 2r}, \quad G = \begin{bmatrix}\tilde{G}_r & -\tilde{G}_i\\ \tilde{G}_i & \tilde{G}_r\end{bmatrix} \in \mathbb{R}^{2d\times 2d}. \tag{80}$$

We notice that $\tilde{G}$ is Hermitian and, therefore, G is symmetric, with its component $\tilde{G}_i$ being skew-symmetric.

Separating the real and imaginary components of $\tilde{G}$, we obtain:

$$\begin{aligned}\tilde{G} &= \tilde{M}\tilde{C}\tilde{M}^H = (\tilde{M}_r + j\tilde{M}_i)\,\tilde{C}\,(\tilde{M}_r + j\tilde{M}_i)^H \\ &= (\tilde{M}_r\tilde{C} + j\tilde{M}_i\tilde{C})(\tilde{M}_r^\intercal - j\tilde{M}_i^\intercal) \\ &= \tilde{M}_r\tilde{C}\tilde{M}_r^T + \tilde{M}_i\tilde{C}\tilde{M}_i^\intercal + j(\tilde{M}_i\tilde{C}\tilde{M}_r^\intercal - \tilde{M}_r\tilde{C}\tilde{M}_i^\intercal).\end{aligned} \tag{81}$$

Representing the parameters of the complex Bingham distribution as in (79), the product between parameter matrices in the real-valued matrix Bingham density can be expressed as:

$$\begin{aligned}\bar{G} &= MCM^\intercal \\ &= \begin{bmatrix}\tilde{M}_r & -\tilde{M}_i \\ \tilde{M}_i & \tilde{M}_r\end{bmatrix}\begin{bmatrix}\tilde{C} & \mathbf{0} \\ \mathbf{0} & \tilde{C}\end{bmatrix}\begin{bmatrix}\tilde{M}_r & \tilde{M}_i \\ -\tilde{M}_i & \tilde{M}_r\end{bmatrix} \\ &= \begin{bmatrix}\tilde{M}_r\tilde{C} & -\tilde{M}_i\tilde{C} \\ \tilde{M}_i\tilde{C} & \tilde{M}_r\tilde{C}\end{bmatrix}\begin{bmatrix}\tilde{M}_r & \tilde{M}_i \\ -\tilde{M}_i & \tilde{M}_r\end{bmatrix} \\ &= \begin{bmatrix}\tilde{M}_r\tilde{C}\tilde{M}_r^T + \tilde{M}_i\tilde{C}\tilde{M}_i^\intercal & \tilde{M}_r\tilde{C}\tilde{M}_i^\intercal - \tilde{M}_i\tilde{C}\tilde{M}_r^\intercal \\ \tilde{M}_i\tilde{C}\tilde{M}_r^\intercal - \tilde{M}_r\tilde{C}\tilde{M}_i^\intercal & \tilde{M}_r\tilde{C}\tilde{M}_r^T + \tilde{M}_i\tilde{C}\tilde{M}_i^\intercal\end{bmatrix}\end{aligned} \tag{82}$$

Comparing this result to (81) and considering the structure for the real parameter matrix G defined in (80), we notice that $\bar{G}$ and G coincide. The parameter matrix G defines a matrix Bingham distribution on $\mathrm{St}(2d, 2r)$ while preserving the properties of multiplications between complex-valued matrices.

Given the structure of the matrices X and G in (80), this real-valued representation of the complex matrix Bingham distribution preserves the invariance properties under right unitary transformations on $\mathrm{U}(d)$, being $p_{\mathcal{B}}(X|M, C)$ invariant under right orthogonal transformations on $\mathrm{O}(2d)$, as noticed in Bingham et al. (1992). Therefore, the derived real-valued matrix Bingham distribution on $\mathrm{St}(2d, 2r)$ is also equivalently defined on the complex Grassmann manifold $\mathrm{Gr}_{\mathbb{C}}(d, r)$. □

K EFFICIENT GEODESICS COMPUTATION

We have obtained an efficient method to compute the KL-divergence between two matrix Bingham distributions leveraging the approximation of the Bingham normalizing constant (32) to reduce the computational complexity of the evaluation procedure. To simplify the notation, we define $p_B(X|\mathbf{z}) = p(X|M(\mathbf{z}), \mathrm{diag}(\mathbf{c}(\mathbf{z})))$, where $M : \mathcal{Z} \to \mathrm{St}(d, k)$ and $\mathbf{c} : \mathcal{Z} \to \mathbb{R}^k_{\geq 0}$ are neural networks representing, respectively, the directional matrix M and diagonal of the concentration matrix C parameterizing the matrix Bingham density. We notice that a neural network providing matrix $M \in \mathrm{St}(d, k)$ can be achieved, e.g., by means of the SVD or QR decomposition of a matrix with the same dimensions of M. In this case, SVD and QR act as retractions on the Stiefel manifold. Both of them are differentiable as proven in (Townsend, 2016) and (Roberts and Roberts, 2020), respectively.

As shown in (Arvanitidis et al., 2022), the knowledge of the KL-divergence between two points of the output statistical manifold leads to an efficient method to compute approximate geodesics by minimizing the energy

$$\mathrm{E}(c) = \sum_{n=1}^{N-1} D_{\mathrm{KL}}(p(\mathbf{x}|c(t_n)) \parallel p(\mathbf{x}|c(t_{n+1})))\,\mathrm{d}t. \tag{83}$$

We can specialize (83) to the case of two matrix Bingham distributions as:

$$\mathrm{E}(c) = \sum_{n=1}^{N-1} \tilde{D}_{\mathrm{KL}}(p_B(X|c(t_n)) \parallel p_B(X|c(t_{n+1}))\,\mathrm{d}t, \tag{84}$$

where $\tilde{D}_{\mathrm{KL}}$ is defined in (62) using the density normalizing constant approximation for the matrix Bingham distribution as in (32).

We implemented the optimization of latent geodesics leveraging the Stochman library (Detlefsen et al., 2021), which we integrated with the theoretical tools developed in this paper.

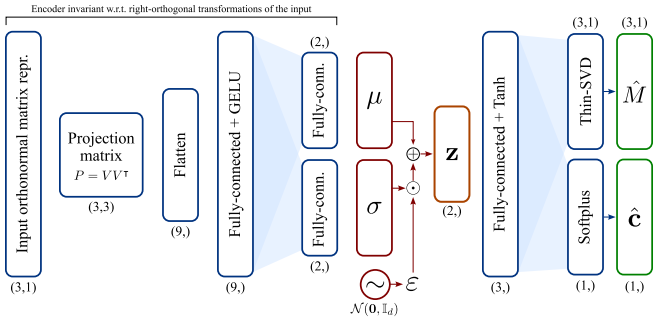


Figure 7: VAE architecture used for the simple synthetic experiment. The network takes as input a unitary vector (which lies on $\mathrm{Gr}(3,1)$) representing the input subspace and outputs the directional and concentration parameters of a matrix Bingham pdf. The tuples under the layers depict the number of output nodes and layer shapes.

L GROUP-INVARIANT SUBSPACE REPRESENTATIONS

The orthogonal projection matrices onto an r-dimensional subspace (or its upper triangular values, being symmetric) bijectively represent the set of subspaces of dimension r.

Proposition L.1. *Let $\mathcal{W} \in Gr(d,r)$ be an r-dimensional subspace, let $(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r)$ be an orthonormal basis for $\mathcal{W}$, and let $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$ be a matrix whose columns are the basis vectors. Then, the mapping*

$$\begin{aligned} \Pi : Gr(d,r) &\to \mathbb{R}^{d\times d}, \\ \Pi(span(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r)) &\mapsto VV^T \end{aligned} \tag{85}$$

is a bijection.

Since the columns of V span the subspace $\mathcal{W} \in \mathrm{Gr}(d,r)$ and $V \in \mathrm{St}(d,r)$ by hypothesis, $P = VV^T \in \mathbb{R}^{d\times d}$ is the orthogonal projection matrix into $\mathcal{W}$. As remarked in (6), there exists an isomorphism between the Grassmann manifold $\mathrm{Gr}(d,r)$ and the set of orthogonal projection matrices of rank r in $\mathbb{R}^{d\times d}$, which makes the mapping Π a bijection between r-dimensional linear subspaces of $\mathbb{R}^d$ and orthogonal projection matrices into those subspaces. Therefore, this representation bijectively maps a subspace $\mathcal{W} \in \mathrm{Gr}(d,r)$ into coordinates representations in $\mathbb{R}^{d\cdot(d+1)/2}$.

On the other hand, a $O(r)$-invariant encoding architecture can be obtained by defining a sequence of layers that takes as input a representation on $\mathrm{St}(d,r)$ and output a set of features that do not vary when a right-orthogonal transformation of the input is performed. An example of such a neural network architecture is given by the GrNet model introduced in (Huang et al., 2018). An input representation on $\mathrm{St}(d,r)$ involves a lower number of parameters with respect to a projection matrix, since its size scales linearly with the space dimension d, against the quadratic scaling of projection matrices.

M EXPERIMENT ON SYNTHETIC DATA

We implemented the VAE models in PyTorch according to the layer types and hyperparameters specified in Fig. 7 for the synthetic experiment. The model is trained by first optimizing the $\hat{M}$ directional component and keeping the concentration fixed at 50. Then, the model is trained including also the predicted concentration in the computation of the matrix Bingham negative log-likelihood used in the VAE loss. The model has been optimized using Adam with an initial learning rate $\gamma = 5 \cdot 10^{-3}$.

In this experiment, we extrapolated the diagonal values of the concentration matrix C towards 0 when this happens. As proposed in (Arvanitidis et al., 2018) for Gaussian decoders, this can be achieved by using a positive radial basis function network to output the diagonal values of C and K-means clustering on the set of training latent codes to determine whether an inferred latent point is near or far from the latent data support. We provide further details on this step in Appendix A.1.

In both cases with and without uncertainty modeling, the geodesics have been determined by discretizing the latent manifold on a grid and computing approximate discrete geodesics curves using the Dijkstra algorithm.

We refined the obtained curves by fitting cubic splines and minimizing the energy (83). When uncertainty is not expressly modeled, the learned manifold appears to be almost flat and, as expected, it does not produce meaningful geodesics that traverse the data domain.

N EXPERIMENTS ON SPACE-TIME CHANNEL SUBSPACES

In this appendix, we detail the experiments on low-dimensional space-time (ST) wireless channel subspaces. In wireless communications, the radio propagation environment between a transmitter (Tx) and a receiver (Rx) is routinely estimated in terms of a channel impulse response matrix. This is a fundamental step in the current communications standard to compensate at the receiver for the distortions produced by the environment on the transmitted signal. At high frequencies, the radio-propagation channel matrices are known to span low-dimensional space-time (ST) subspaces. Such subspaces can be estimated using low-rank estimation over multiple local channel estimates efficiently obtained, e.g., by Least Squares (LS) estimation using a set of transmitted symbols (called pilots) known at the receiver.

The aims of this experiment are to:

- Auto-encode the ST subspaces of the wireless channel using a VAE that takes as input the orthonormal bases representing the Tx/Rx ST subspaces and outputs the parameters of the matrix Bingham distributions of the reconstructed subspaces.
- Model the uncertainty on the VAE output concentration parameters using radial basis functions accounting for the distance between the predicted latent codes and the training ones, as proposed in Arvanitidis et al. (2022).
- Pull-back the Fisher-Rao metrics of the output matrix Bingham distributions into the latent space to achieve meaningful and identifiable latent representations.

In the following, we detail the data generation process, the adopted VAE architecture, the training procedure and the results.

N.1 System model

In this experiment, we considered a narrow-bandwidth uplink multiple-input multiple-output vehicle-to-infrastructure (V2I) communication system over a bandwidth B featured by a base station (BS) and a set of vehicular equipment (VEs) crossing an urban setting. The BS is equipped with an $N_R^{az} \times N_R^{el}$ rectangular antenna array while the VEs are equipped with $N_T^{az} \times N_T^{el}$ rectangular antenna arrays, where az denotes the azimuth direction and el represents the elevation direction of the antenna array. At the receiver, the discrete-time Rx signal is modeled, after synchronizing in time and frequency and removing the cyclic prefix, as:

$$\mathbf{y}[w] = \mathbf{H}[w] * \mathbf{x}[w] + \mathbf{n}[w], \tag{86}$$

for $w \in 1, \cdots, W$, where W is the maximum number of temporal channel taps, and the discrete-time system has been sampled at time $t = wT$, $T = 1/B$. $\mathbf{y}[w] \in \mathbb{C}^{N_R \times 1}$ is the received signal, $\mathbf{x}[w] \in \mathbb{C}^{N_T \times 1}$ is the transmitted signal, $\mathbf{H}[w] \in \mathbb{C}^{N_R \times N_T}$ denotes the MIMO channel matrix, and $\mathbf{n}[w] \in \mathbb{C}^{N_R \times 1}$ is additive Gaussian noise that corrupts the Rx signal. In the following, we will employ the channel matrix in the form $\mathbf{h} = \mathrm{vec}\left[\mathrm{vec}(\mathbf{H}[1]) \cdots \mathrm{vec}(\mathbf{H}[W])\right] \in \mathbb{C}^{WN_TN_R \times 1}$, which aggregates the channel impulse responses over the channel taps in a single vector.

Table 1 reports the communication system parameters used for the simulations.

N.1.1 Channel model

At high frequencies, the MIMO channel between the Tx and the Rx can be routinely modeled as the sum of few propagation paths (Akdeniz et al., 2014):

$$\mathbf{H}[w] = \sum_{p=1}^{P} \beta_p \, e^{j2\pi\nu_p wT} \, \mathbf{a}_R(\boldsymbol{\theta}_p)\mathbf{a}_T^{\mathrm{T}}(\boldsymbol{\phi}_p) \, g\left[wT - \tau_p\right], \tag{87}$$

Table 1: Channel simulation parameters. For the interaction types, R stands for reflection, D for diffraction and DS for diffuse scattering.

Simulation parameter	Value
Carrier frequency	28 GHz
Bandwidth	1 MHz
BS antenna array size (az. $\times$ el.)	8×8
VE antenna array size (az. $\times$ el.)	4×4
BS height from the ground	7 m
VE height from the ground	1.67 m – 4.41 m
SUMO simulation sampling time	0.1 s
Sionna Number of initial sampled rays	10^6
Sionna ray tracing method	Fibonacci
Selected interaction types	R, D, DS
Signal-to-noise ratio (SNR)	-15 dB

where

- β_p is the p-th path amplitude and depends on path-loss and on the geometry of the propagation environment.
- ν_p is the p-th path Doppler shift.
- $\mathbf{a}_T(\boldsymbol{\phi}_p) \in \mathbb{C}^{N_T \times 1}$ and $\mathbf{a}_R(\boldsymbol{\theta}_p) \in \mathbb{C}^{N_R \times 1}$ are the Tx and Rx array response vectors to the p-th path; they are function of the directions of the departure (DoDs) $\boldsymbol{\phi}_p = [\phi_p^{\text{az}}, \phi_p^{\text{el}}]^{\text{T}}$ and the directions of arrival (DoAs) $\boldsymbol{\theta}_p = [\theta_p^{\text{az}}, \theta_p^{\text{el}}]^{\text{T}}$ (for azimuth and elevation).
- $g(wT - \tau_p)$ is the sampled pulse shaping waveform (typically a raised cosine) delayed by τ_p (p-th path delay).

We consider uniform planar arrays composed by isotropic antennas spaced at half-wavelength for both Tx and Rx. The response of the Tx antenna array is given by:

$$\mathbf{a}_T(\boldsymbol{\phi}_p) = \mathbf{a}_T^{\text{el}}(\phi_p^{\text{el}}) \otimes \mathbf{a}_T^{\text{az}}(\phi_p^{\text{az}}), \tag{88}$$

$$\mathbf{a}_T^{\text{az}}(\phi_p^{\text{az}}) = [1, \dots, e^{j\pi(N_T-1)\sin(\phi_p^{\text{az}})}] \tag{89}$$

$$\mathbf{a}_T^{\text{el}}(\phi_p^{\text{el}}) = [1, \dots, e^{j\pi(N_T-1)\sin(\phi_p^{\text{el}})}] \tag{90}$$

where $\mathbf{a}_T^{\text{az}}(\phi_p^{\text{az}})$ and $\mathbf{a}_T^{\text{el}}(\phi_p^{\text{el}})$ are the steering vectors along the array azimuth and elevation directions. The array response at the Rx is similarly defined. For more details on the system and channel models, we refer the reader to (Brighente et al., 2020; Cazzella et al., 2022).

N.2 Data generation

We detail here the generation of the subspace data used to perform subspace auto-encoding and pull-back of the matrix Bingham information geometry over space-time wireless channel subspaces.

Maps and roads information retrieval: To simulate the wireless communication channel, we retrieved from OpenStreetMap (OSM)[3] the roads geometry and the 3D meshes of the buildings featuring an urban area depicted in Fig. 8, where the BS is indicated by the red circle.

Vehicular traffic simulation: Besides the 3D meshes data, we retrieved from OSM also the information on the roads geometry. We imported these details into the SUMO vehicular traffic simulator (Lopez et al., 2018), which we used to generate realistic traffic data on the chosen urban setting. SUMO is an efficient microscopic simulator, i.e., it models the vehicular traffic flow considering the single vehicles as independent entities. The selected roads are depicted in Fig. 8, where the simulated trajectories are represented using a 2D color gradient to ease the visualization of their correspondence with the latent points learned by the considered latent variable model.

[3] https://www.openstreetmap.org/

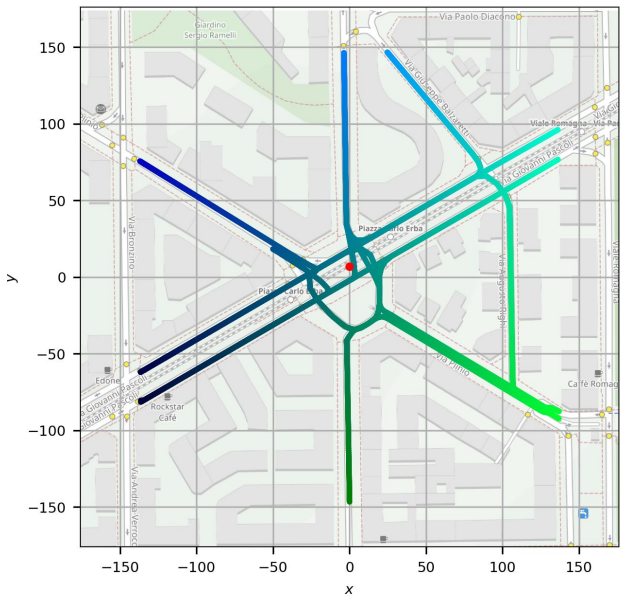


Figure 8: Top view of the considered urban vehicular scenario. The colored lines depict the simulated vehicular trajectories, which have been considered for the generation of wireless channel matrices. The red point represents the position of the BS.

Ray-tracing wireless channel simulation: We used the NVIDIA Sionna RT (Hoydis et al., 2023) ray-tracing engine to simulate the communication channel between the BS and VEs. Sionna RT allowed us to accurately simulate high frequency radio propagation by modeling reflection, diffraction and diffuse scattering interactions with the environment. Referring to Fig. 8, two antenna arrays pointing towards opposite directions on the y-axis have been considered at the BS according to the parameters reported in Table 1. The antenna arrays are parallel to the xz-plane, where the z-axis is orthogonal to the image plane.

Low-Rank channel estimation: In wireless vehicular communications, the radio-propagation channels are estimated in matrix form and are known to span low-dimensional space-time (ST) subspaces. Such subspaces can be estimated using low-rank estimation over multiple channel samples (obtained, e.g., by Least Squares). We adopted the multi-vehicular estimation procedure described in (Cazzella et al., 2022; Mizmizi et al., 2021) to generate the spatial channel subspaces owing to the transmitter (Tx) and the receiver (Rx). This procedure exploits multiple passages of vehicular communication equipment (VEs) in the same areas to compute the space-time subspaces spanned by the channel. For the analytical details of LR channel estimation, we refer to (Brighente et al., 2020).

The Sionna RT and SUMO simulation parameters are reported in Table 1. The training dataset used in this experiment is composed of a set of ST subspaces estimated through low-rank estimation from the simulated wireless channels. The latter are generated between each VE and the BS.

N.3 VAE architecture

The VAE architecture considered in this experiment is depicted in Fig. 9. Since we considered a narrow-band communication system, the temporal component is not relevant and, therefore, is not considered. We separately model the spatial channel subspace owing to the Tx and the one owing to the Rx. For each data point, the subspaces are represented by the complex-valued orthonormal matrices $\bar{U}_{\text{Tx}} \in \text{St}_{\mathbb{C}}(64, 3)$ and $\bar{U}_{\text{Rx}} \in \text{St}_{\mathbb{C}}(16, 2)$, spanning, respectively, the spatial subspace at the Tx and the spatial subspace at the Rx. Using the procedure discussed in Appendix J, we computed the real-valued orthonormal representations $U_{\text{Tx}} \in \text{St}(128, 6)$ and $U_{\text{Rx}} \in \text{St}(32, 4)$, which are equivalent to the complex-valued ones under matrix sum and multiplication.

We provided as input to the network the orthonormal matrices U_{Tx} and U_{Rx}. Both matrices are processed by two parallel encoders that are invariant to right-orthogonal transformations of the inputs, and, therefore, consistently

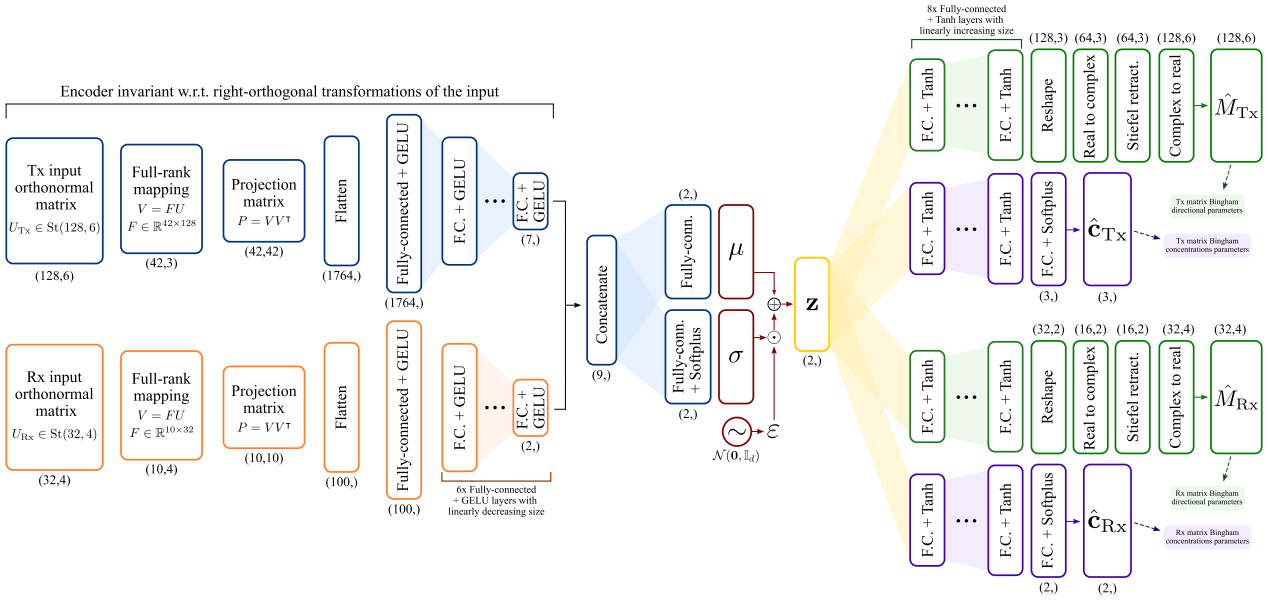


Figure 9: VAE architecture used for the experiment on space-time channel subspaces. The network takes as inputs two orthonormal bases representations on $\text{St}(128, 6)$ and $\text{St}(32, 4)$, respectively, corresponding to the spatial subspace owing to the transmitter (Tx) and the receiver (Rx). The tuples under the layers depict the number of output nodes and shapes.

model the subspaces spanned by the input orthonormal matrices rather than the matrices themselves. The encoding outputs are concatenated and the resulting vector is mapped through two fully-connected layers to the mean $\mu \in \mathbb{R}^2$ and standard deviation $\sigma \in \mathbb{R}^2_+$, used to model a latent Gaussian distribution with diagonal covariance matrix $\mathcal{N}(\mu, \text{diag}(\boldsymbol{\sigma^2}))$. We employed a Softplus activation for the standard deviation estimation layer to ensure that the obtained std. dev. is positive.

After sampling, the latent codes are mapped to 4 fully-connected decoders with Tanh activations. The VAE's output is provided by the directional and concentration parameters of two real-valued matrix Bingham pdfs: $p_{\mathcal{B}}(U_{\text{Tx}}|\hat{M}_{\text{Tx}}, \hat{\mathbf{c}}_{\text{Tx}})$ and $p_{\mathcal{B}}(U_{\text{Rx}}|\hat{M}_{\text{Rx}}, \hat{\mathbf{c}}_{\text{Rx}})$. Two branches allow for the estimation of the directional parameters $\hat{M}_{\text{Tx}}$ and $\hat{M}_{\text{Rx}}$ by:

1. Reshaping the fully-connected layers output.

2. Equally splitting the resulting matrix rows into the real and imaginary parts of a complex-valued vector and computing the resulting complex vector.

3. Retracting the obtained matrix on the complex Stiefel manifolds $\text{St}_{\mathbb{C}}(64, 3)$ and $\text{St}_{\mathbb{C}}(16, 2)$, respectively.

4. Mapping the obtained complex-valued orthonormal matrix into the corresponding real-valued representation as in (80).

We notice that the two modeled real-valued matrix Bingham distributions are equivalent to two complex-valued matrix Bingham distributions according to the mapping discussed in Appendix J. Therefore, they can model, without loss of generality, the distribution of the spatial Tx and Rx subspaces on the considered scenario.

In this experiment, we extrapolated the diagonal values of the concentration matrix C towards 0 when this happens. As proposed in (Arvanitidis et al., 2018) for Gaussian decoders, this can be achieved by using a positive radial basis function network to output the diagonal values of C and K-means clustering on the set of training latent codes to determine whether an inferred latent point is near or far from the latent data support. We provide further details on this step in Appendix A.1.

N.4 Training procedure

The model has been trained by minimizing the modified ELBO loss:

$$\begin{aligned}\mathcal{L} = \sum_{i=1}^{N} \frac{1}{2} \Big(&- \log p_{\theta,\mathrm{Tx}}(U_{\mathrm{Tx},i}|\hat{M}_{\mathrm{Tx}}(\mathbf{z}_i), \hat{\mathbf{c}}_{\mathrm{Tx}}(\mathbf{z}_i)) - \log p_{\theta,\mathrm{Rx}}(U_{\mathrm{Rx},i}|\hat{M}_{\mathrm{Rx}}(\mathbf{z}_i), \hat{\mathbf{c}}_{\mathrm{Rx}}(\mathbf{z}_i)) \Big) \\ &+ \beta \cdot KL\left(\mathcal{N}(\mathbf{z}_i|\mu_\phi(U_{\mathrm{Tx},i}, U_{\mathrm{Rx},i}), \mathrm{diag}(\sigma^2(U_{\mathrm{Tx},i}, U_{\mathrm{Rx},i}))) \parallel \mathcal{N}(\mathbf{0}, \mathbb{I}_d)\right)\end{aligned} \tag{91}$$

where N is the cardinality of the training dataset, and we used mini-batches of size 64 to minimize the loss stochastically through the Adam optimizer Kingma (2014). $p_{\theta,\mathrm{Tx}}(U_{\mathrm{Tx},i}|\hat{M}_{\mathrm{Tx}}(\mathbf{z}_i), \hat{\mathbf{c}}_{\mathrm{Tx}}(\mathbf{z}_i))$ and $p_{\theta,\mathrm{Rx}}(U_{\mathrm{Rx},i}|\hat{M}_{\mathrm{Rx}}(\mathbf{z}_i), \hat{\mathbf{c}}_{\mathrm{Rx}}(\mathbf{z}_i))$ are the pdfs of two matrix Bingham distributions as in (9). The first term in the loss represents the average of the negative log-likelihoods for the estimated matrix Bingham parameters on the Tx and Rx spatial subspaces, while $KL(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence between two pdfs. The distribution parameters have been indicated in functional form to make explicit their dependency on the inputs or the latent variable $\mathbf{z}$.

β is a hyperparameter controlling the weighting of the KL component with respect to the negative log-likelihood one. During training, we employed a KL annealing schedule for β so that, depending on the epoch the β starts from 0 and gradually increases up to 1 according to the following function:

$$\beta_k = \mathrm{Tanh}(a \cdot k) \tag{92}$$

where the subscript $k = 1, \ldots, K$ denotes the training epoch and we set the annealing factor $a = 10^{-3}$.

The training procedure has been divided into two optimization problems, which we observed that allowed to obtain a higher performance:

1. First, the VAE decoder's parameters related to the two branches estimating the Tx and Rx matrix Bingham concentrations have been frozen, optimizing only the directional parameters M_{Tx} and M_{Rx} while fixing all the distributions' concentrations to 20.
2. As a second step, we jointly optimized both the directional and concentration parameters. This allowed us training the model to predict meaningful concentrations $\mathbf{c}_{\mathrm{Tx}}$ and $\mathbf{c}_{\mathrm{Rx}}$ near the data support.

The dataset has been randomly split by vehicular trajectory into training and validation sets (80% training; 20% validation) to assess the generalization capabilities of the model to input data unseen at training time.

N.5 Latent Riemannian pull-back metric computation

We determined the latent metric as the average of the pulled-back Fisher-Rao metrics induced by the output matrix Bingham distributions modeling the Tx and Rx spatial channel subspaces.

N.6 Experimental results

We trained the VAE model using the hyperparameters reported in Table 2. To assess the reconstruction capabilities of the VAE, we used the estimated subspaces to filter a set of Least Squares (LS) channel estimates generated for the same VE locations as for the input channel subspaces. Therefore, we started from a set of LS channel estimates $\mathbf{h}_{\mathrm{LS},i} \in \mathbb{C}^{N_T N_R \times 1}$, $i \in 1, \ldots, N$ and associated Tx/Rx spatial subspaces represented by the orthonormal bases $U_{\mathrm{Tx},i} \in \mathbb{C}^{N_T \times r_s^{\mathrm{Tx}}}$ and $U_{\mathrm{Rx},i} \in \mathbb{C}^{N_R \times r_s^{\mathrm{Rx}}}$. We performed low-rank channel estimation using the subspaces estimated by the VAE, and we measured the normalized mean squared error (NMSE) of the LS and LR estimates as:

$$\mathrm{NMSE} = \frac{\mathbb{E}\left[\|\hat{\mathbf{h}} - \mathbf{h}\|_2^2\right]}{\mathbb{E}\left[\|\mathbf{h}\|_2^2\right]}, \tag{93}$$

where $\hat{\mathbf{h}}$ is a channel estimate, either LS or LR through filtering by the subspaces provided by the considered VAE model, and $\mathbf{h}$ is the ground truth simulated channel impulse response matrix. $\hat{\mathbf{h}}_{\mathrm{LR}}$ is provided by:

$$\hat{\mathbf{h}}_{\mathrm{LR},i} = \hat{\Pi}_{\mathrm{LR},i} \mathbf{h}_{\mathrm{LS},i}, \tag{94}$$

Table 2: Hyperparameters selected for the ST channel subspaces experiment.

Hyperparameter	Symbol	Value
Latent space dim.	d	2
Bingham norm. const. truncation	γ_{nc}	3
Bingham direction training epochs num.	K_d	1500
Bingham concentration training epochs num.	K_c	250
Tx spatial subspaces dim.	r_s^{Tx}	3
Rx spatial subspaces dim.	r_s^{Rx}	2
Initial learning rate	ν	10^{-4}
K-means centroids num. (RBF u.q.)	C	200
RBF u.q. c_{RBF}	c_{RBF}	5
RBF u.q. β_{RBF}	β_{RBF}	-3.2

with

$$\hat{\Pi}_{\text{Tx},i} = \hat{U}_{\text{Tx},i}\hat{U}_{\text{Tx},i}^H, \tag{95}$$

$$\hat{\Pi}_{\text{Rx},i} = \hat{U}_{\text{Rx},i}\hat{U}_{\text{Rx},i}^H, \tag{96}$$

$$\hat{\Pi}_{\text{LR},i} = \hat{\Pi}_{\text{Tx},i}^* \otimes \hat{\Pi}_{\text{Rx},i}, \tag{97}$$

and we considered the modes of the matrix Bingham distributions as the spatial Tx and spatial Rx subspace estimates provided by the VAE. We notice that the mode of a matrix Bingham is provided by the leading eigenvectors of its directional parameter M, as detailed in Appendix I.

N.7 Comparison with subspace regression

To reach a more direct comparison with respect to previous methods, we have considered the model proposed in Cazzella et al. (2022) and we have performed a set of experiments to compare it with our model on the high-frequency EM propagation setup detailed in Section 5.2 of the main paper. Even if the two model architectures are not compatible for direct comparison, we targeted the end-to-end auto-encoding performance in terms of normalized mean squared error (NMSE) as comparison metric. This allowed us to highlight that both methods attain the ground truth reference NMSE performance by a low margin, with our method additionally providing latent geometrical modeling and uncertainty quantification over the predicted subspaces.

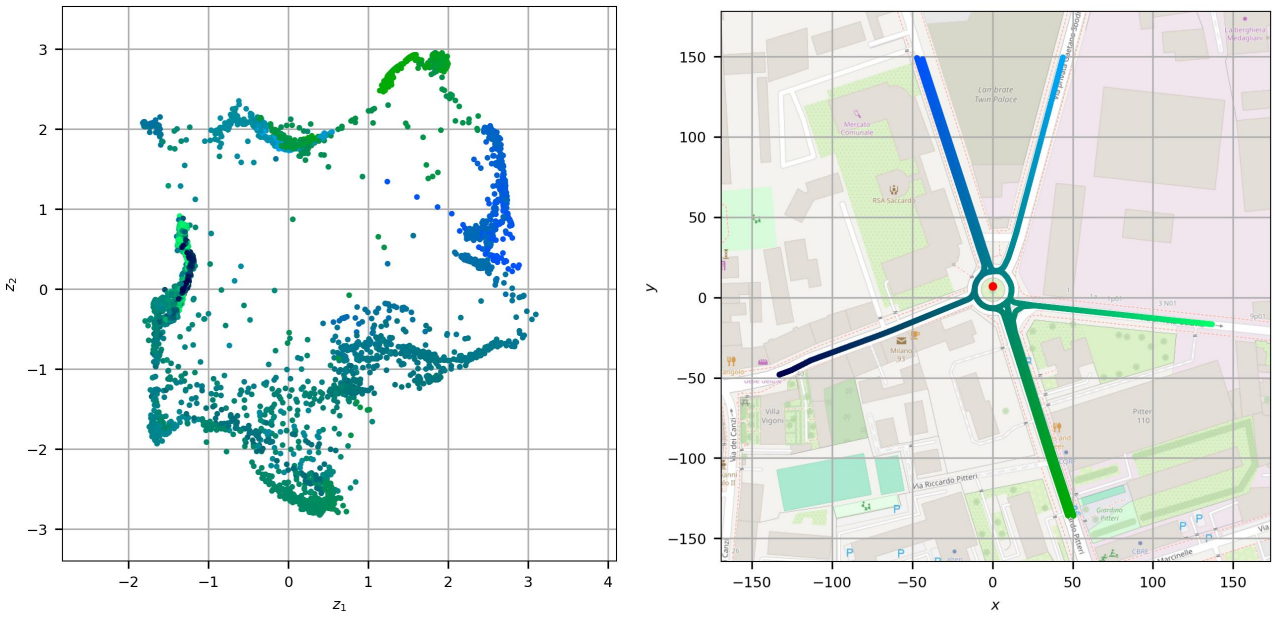
To the best of our knowledge, this work is the first to introduce a VAE architecture for subspace data, which prevents a direct comparison with other models. The model proposed in (Cazzella et al., 2022) does not address the variational auto-encoding or generative modeling over subspace ambient spaces, targeting the regression of the subspaces on which lie the input noisy channel estimates.

N.8 Evaluation across different conditions

We detail here the experiments performed over conditions different from the ones of the main scenario. In particular, we considered different SNR, randomization seeds and EM propagation setting to assess the model's performance. The following experiments are based on variations of the setup examined in Sec. N.6.

Different SNR. We considered a different signal-to-noise ratio (SNR) at -30 dB for the wireless communications scenario examined above. Also in this case, we obtained a significant NMSE ratio (-22.06 dB on the validation set). We notice that the quality of the subspaces used during training does not change considerably with respect to the first scenario since they are determined using a high number of channel estimates, which has been shown to converge to high quality subspace estimates in (Mizmizi et al., 2021). What changes is the noisiness of the channel estimates that are projected on the spatial subspaces predicted by the model. The aim of these experiments is to prove the feasibility of this method and to show that it can denoise estimates at different SNRs. Another possible line of experiments, which we do not target in this paper and we consider as prospective work, can involve the study of the auto-encoding performance under noisy subspaces.

Different randomization seeds. We considered experiments on a set of 5 different seeds for the same wireless communications scenario discussed in the paper, obtaining a mean NMSE gain of -18.24 dB over the range



(a) Latent space (mean)

(b) Roads geometry

Figure 10: a) Latent representations (means) for the training data points obtained for the high-frequency EM propagation experiment on the second scenario. b) Top view of the considered alternative urban vehicular scenario. The roads' 2D color gradient is mapped to the colors of the latent points. The red point represents the BS.

(-19.6 dB, -17.61 dB), along with well-structured latent spaces as for the experiments proposed in the paper, showing that the estimated subspaces perform consistently on the low-rank channel estimation task.

Different EM propagation scenario. To provide greater evidence for the generalization of the proposed methodology to different wireless communications conditions, we have selected the wireless communications scenario depicted in Fig. 10b. The remaining hyperparameters and channel estimation/processing procedures have been set as for the experiments above. We obtained channel estimation results comparable to the first scenario, with an NMSE gain of -17.4 dB on the validation set, while achieving well-structured latent representations (see Fig. 10a).

O ALTERNATIVE SUBSPACE PROBABILITY DISTRIBUTIONS

Besides the matrix Bingham distribution, we identified the matrix angular central Gaussian (MACG) distribution proposed in (Chikuse, 1990), which we target for further investigation as probability distribution over subspaces. Differently from the matrix Bingham distribution, the MACG distribution has more favorable sampling and optimization properties but involves a higher dimensional parameters space, which make it not directly suitable to model high-dimensional subspace domains that are intrinsically low-rank.

P LLM POLICY

No LLM service has been used in any phase of the development and writing of this work.

Q CODE RELEASE

We release the code covering the techniques developed in this paper in the following public repository: https://github.com/lorenzocazzella/grassmann-latent-information-geometry.