Bayesian Analysis (2020) **15**, Number 3, pp. 871–908

Conjugate Priors and Posterior Inference for the Matrix Langevin Distribution on the Stiefel Manifold

Subhadip Pal*, Subhajit Sengupta†, Riten Mitra‡, and Arunava Banerjee§

Abstract. Directional data emerges in a wide array of applications, ranging from atmospheric sciences to medical imaging. Modeling such data, however, poses unique challenges by virtue of their being constrained to non-Euclidean spaces like manifolds. Here, we present a unified Bayesian framework for inference on the Stiefel manifold using the Matrix Langevin distribution. Specifically, we propose a novel family of conjugate priors and establish a number of theoretical properties relevant to statistical inference. Conjugacy enables translation of these properties to their corresponding posteriors, which we exploit to develop the posterior inference scheme. For the implementation of the posterior computation, including the posterior sampling, we adopt a novel computational procedure for evaluating the hypergeometric function of matrix arguments that appears as normalization constants in the relevant densities.

Keywords: Bayesian inference, conjugate prior, hypergeometric function of matrix argument, matrix Langevin distribution, Stiefel manifold, vectorcardiography.

1 Introduction

Analysis of directional data is a major area of investigation in statistics. Directional data range from unit vectors in the simplest case to sets of ordered orthonormal frames in the general scenario. Since the associated sample space is non-Euclidean, standard statistical methods developed for the Euclidean space may not be appropriate to analyze such data. Additionally, it is often desirable to design statistical methods that take into consideration the underlying geometric structure of the sample space. There is a need for methodological development for a general sample space such as the Stiefel manifold (James, 1976; Chikuse, 2012) that goes beyond those techniques designed for simpler non-Euclidean spaces like the circle or the sphere. Such a novel methodology can support various emerging applications, increasingly seen in the fields of biology (Downs, 1972; Mardia and Khatri, 1977), computer science (Turaga et al., 2008; Lui and Beveridge, 2008) and astronomy (Mardia and Jupp, 2009; Lin et al., 2017), to mention but a few.

*Department of Bioinformatics and Biostatistics, University of Louisville, subhadip.pal@louisville.edu
†Center for Psychiatric Genetics, NorthShore University HealthSystem, ssengupta@northshore.org
‡Department of Bioinformatics and Biostatistics, University of Louisville, r0mitr01@louisville.edu
§Department of Computer & Information Science & Engineering, University of Florida, arunava@ufl.edu

 https://doi.org/10.1214/19-BA1176

One of the most widely used probability distributions on the Stiefel manifold is the matrix Langevin distribution introduced by Downs (1972), also known as the Von-Mises Fisher matrix distribution (Mardia and Jupp, 2009; Khatri and Mardia, 1977). In early work, Mardia and Khatri (1977) and Jupp and Mardia (1980) investigated properties of the matrix Langevin distribution and developed inference procedures in the frequentist setup (Chikuse, 2012). The form of the maximum likelihood estimators and the profile likelihood estimators for the related parameters can be found in Jupp and Mardia (1979); Mardia and Khatri (1977); Chikuse (1991b,a, 1998). It is not patently clear from these works whether the form of the associated asymptotic variance can be obtained directly without using bootstrap procedures. A major obstacle facing the development of efficient inference techniques for this family of distributions has been the intractability of the corresponding normalizing constant, a hypergeometric function of a matrix argument (Mardia and Jupp, 2009; Muirhead, 2009; Gross and Richards, 1989). Inference procedures have been developed exploiting approximations that are available when the argument to this function is either small or large.

Almost all the hypothesis testing procedures (Jupp and Mardia, 1979; Mardia and Khatri, 1977; Chikuse, 1991a,b, 1998) therefore depend not only on large sample asymptotic distributions but also on the specific cases when the concentration parameter is either large or small (Chikuse, 2012; Mardia and Khatri, 1977; Downs, 1972). In particular, a general one sample or two sample hypothesis testing method for the finite sample case is yet to be developed.

For any given dataset, the stipulation of large sample is comparatively easier to verify than checking whether the magnitude of the concentration is large. It may not be possible to ascertain whether the concentration is large before the parameter estimation procedure, which is then confounded by the fact that the existing parameter estimation procedures themselves require the assumption of large concentration to work correctly. Hence, from a practitioner's point of view, it is often difficult to identify whether the above-mentioned procedures are suitable for use on a particular dataset.

Although a couple of Bayesian procedures have been proposed in related fields (see references in Lin et al. (2017)), a comprehensive Bayesian analysis is yet to be developed for the matrix Langevin distribution. In a recent paper, Lin et al. (2017) have developed a Bayesian mixture model of matrix Langevin distributions for clustering on the Stiefel manifold, where they have used a prior structure that does not have conjugacy. To accomplish posterior inference, Lin et al. (2017) have used a nontrivial data augmentation strategy based on a rejection sampling technique laid out in Rao et al. (2016). It is worthwhile to note that the specific type of data augmentation has been introduced to tackle the intractability of the hypergeometric function of a matrix argument. It is well known that data augmentation procedures often suffer from slow rate of convergence (van Dyk and Meng, 2001; Hobert et al., 2011), particularly when combined with an inefficient rejection sampler. Elsewhere, Hornik and Grün (2014) have proposed a class of conjugate priors but have not presented an inference procedure for the resulting posterior distributions.

In this article, we develop a comprehensive Bayesian framework for the matrix Langevin distribution, starting with the construction of a flexible class of conjugate priors, and proceeding all the way to the design of an practicable posterior computation

procedure. The difficulties arising from the intractability of the normalizing constant do not, of course, disappear with the mere adoption of a Bayesian approach. We employ nontrivial strategies to derive a unique posterior inference scheme in order to handle the intractability of the normalizing constant. A key step in the proposed posterior computation is the evaluation of the hyper-geometric function of a matrix argument, that can be computed using the algorithm developed in Koev and Edelman (2006). Although general, this algorithm has certain limitations vis-à-vis measuring the precision of its output. We therefore construct a reliable and computationally efficient procedure to compute a specific case of the hypergeometric function of matrix argument, that has theoretical precision guarantees (Section 6.2). The procedure is applicable to a broad class of datasets including most, if not all, of the applications found in Downs et al. (1971); Downs (1972); Jupp and Mardia (1979, 1980); Mardia and Khatri (1977); Mardia et al. (2007); Mardia and Jupp (2009); Chikuse (1991a,b, 1998, 2003); Sei et al. (2013); Lin et al. (2017). The theoretical framework proposed in this article is applicable to all matrix arguments regardless of dimensionality. In the following two paragraphs, we summarize our contributions.

We begin by adopting a suitable representation of the hypergeometric function of a matrix argument to view it as a function of a vector argument. We explore several of its properties that are useful for subsequent theoretical development, and also adopt an alternative parametrization of the matrix Langevin distribution so that the modified representation of the hypergeometric function can be used. When viewed as an exponential family of distributions, the new parameters of the matrix Langevin distribution are not the natural parameters (Casella and Berger, 2002). Thus the construction of the conjugate prior does not directly follow from Diaconis and Ylvisaker (1979) (DY), an issue that we elaborate on (Section 3.1). We then propose two novel and reasonably large classes of conjugate priors, and based on theoretical properties of the matrix Langevin distribution and the hypergeometric function, we establish their propriety. We study useful properties of the constructed class of distributions to demonstrate that the hyperparameters related to the class of distributions have natural interpretations. Specifically, the class of constructed distributions is characterized by two hyperparameters, one controls the *location* of the distribution while the other determines the *scale*. This interpretation not only helps us understand the nature of the class of distributions but also aids in the selection of hyperparameter settings. The constructed class of prior distributions is flexible because one can incorporate prior knowledge via appropriate hyperparameter selection; and at the same time, in the absence of prior knowledge, there is a provision to specify the hyperparameters to construct a uniform prior. Since this uniform prior is improper by nature, we extend our investigation to identify the conditions under which the resulting posterior is a proper probability distribution.

Following this, we discuss properties of the posterior and inference. We show unimodality of the resulting posterior distributions and derive a computationally efficient expression for the posterior mode. We also demonstrate that the posterior mode is a consistent estimator of the related parameters. We develop a Gibbs sampling algorithm to sample from the resulting posterior distribution. One of the conditionals in the Gibbs sampling algorithm is a novel class of distributions that we have introduced in this article for the first time. We develop and make use of properties such as unimodality and log-concavity to derive a rejection sampler to sample from this distribution. We

perform multiple simulations to showcase the generic nature of our framework and to report estimation efficiency for the different algorithms. We end with an application demonstrating the strength of our approach.

We should note that a significant portion of the article is devoted to establishing a number of novel properties of the hypergeometric function of matrix arguments. These properties play a key role in the rigorous development of the statistical procedures. These properties, including the exponential type upper and lower bounds for the function, may also be relevant to a broader range of scientific disciplines.

The remainder of the article is organized as follows. In Section 2, we introduce the matrix Langevin distribution defined on the Stiefel manifold and explore some of its important properties. Section 3 begins with a discussion of the inapplicability of DY's theorem, following which we present the construction of the conjugate prior for the parameters of the matrix Langevin distribution. In particular, we establish propriety of a class of posterior and prior distributions by proving the finiteness of the integral of specific density kernels. In Sections 4 and 5, we lay out the hyperparameter selection procedure and derive properties of the posterior. In Section 6 we develop the posterior inference scheme. In Sections 7 and 8, we validate the robustness of our framework with experiments using simulated datasets and demonstrate the applicability of the framework using a real dataset, respectively. Finally, in Section 9, we discuss other developments and a few possible directions for future research. Proofs of all theorems and properties of the hypergeometric function of matrix arguments are deferred to the supplementary material (Pal et al., 2019).

Notational Convention

$\mathbb{R}^p$ = The p-dimensional Euclidean space.

$\mathbb{R}^p_+ = \{(x_1, \ldots, x_p) \in \mathbb{R}^p : 0 < x_i \text{ for } i = 1, \ldots p\}$.

$\mathcal{S}_p = \left\{(d_1, \ldots, d_p) \in \mathbb{R}^p_+ : 0 < d_p < \cdots < d_1 < \infty\right\}$.

$\mathbb{R}^{n\times p}$ = Space of all $n \times p$ real-valued matrices.

$\mathbf{I}_p = p \times p$ identity matrix.

$\mathcal{V}_{n,p} = \{X \in \mathbb{R}^{n\times p} \,:\, X^T X = \mathbf{I}_p\}$, Stiefel Manifold of p-frames in $\mathbb{R}^n$.

$\widetilde{\mathcal{V}}_{n,p} = \{X \in \mathcal{V}_{n,p} : X_{1,j} \geq 0 \;\; \forall\, j = 1, 2, \cdots, p\}$.

$\mathcal{V}_{p,p} = O(p)$ = Space of Orthogonal matrices of dimension $p \times p$.

μ = Normalized Haar measure on $\mathcal{V}_{n,p}$.

μ_2 = Normalized Haar measure on $\mathcal{V}_{p,p}$.

μ_1 = Lebesgue measure on $\mathbb{R}^p_+$.

$f(\cdot;\cdot)$ = Probability density function.

$g(\cdot;\cdot)$ = Unnormalized version of the probability density function.

$tr(A)$ = Trace of a square matrix A.

$etr(A)$ = Exponential of $tr(A)$.

$E(X)$ = Expectation of the random variable X.

$\mathbb{I}(\cdot)$ = Indicator function.

$\|\cdot\|_2$ = Matrix operator norm.

We use $\boldsymbol{d}$ and D interchangeably. D is the diagonal matrix with diagonal $\boldsymbol{d}$. We use matrix notation D in the place of $\boldsymbol{d}$ wherever needed, and vector $\boldsymbol{d}$ otherwise.

2 The Matrix Langevin Distribution on the Stiefel Manifold

The Stiefel manifold, $\mathcal{V}_{n,p}$, is the space of all p ordered orthonormal vectors (also known as p-frames) in $\mathbb{R}^n$ (Mardia and Jupp, 2009; Absil et al., 2009; Chikuse, 2012; Edelman et al., 1998; Downs, 1972) and is defined as

$$\mathcal{V}_{n,p} = \{X \in \mathbb{R}^{n\times p} \,:\, X^T X = \mathbf{I}_p,\ p \le n\},$$

where $\mathbb{R}^{n\times p}$ is the space of all $n \times p$, $p \le n$ real-valued matrices, and $\mathbf{I}_p$ is the $p \times p$ identity matrix. $\mathcal{V}_{n,p}$ is a compact Riemannian manifold of dimension $np - p(p+1)/2$ (Chikuse, 2012). A topology on $\mathcal{V}_{n,p}$ can be induced from the topology on $\mathbb{R}^{n\times p}$ as $\mathcal{V}_{n,p}$ is a sub-manifold of $\mathbb{R}^{n\times p}$ (Absil et al., 2009; Edelman et al., 1998). For $p = n$, $\mathcal{V}_{n,p}$ becomes identical to $O(n)$, the orthogonal group consisting of all orthogonal $n \times n$ real-valued matrices, with the group operation being matrix multiplication. Being a compact unimodular group, $O(n)$ has a unique Haar measure that corresponds to a uniform probability measure on $O(n)$ (Chikuse, 2012). Also, through obvious mappings, the Haar measure on $O(n)$ induces a normalized Haar measure on the compact manifolds $\mathcal{V}_{n,p}$. The normalized Haar measures on $O(n)$ and $\mathcal{V}_{n,p}$ are invariant under orthogonal transformations (Chikuse, 2012). Detailed construction of the Haar measure on $\mathcal{V}_{n,p}$ and its properties are described in Muirhead (2009); Chikuse (2012). Notation wise, we will use μ and μ_2 to denote the normalized Haar measures on $\mathcal{V}_{n,p}$ and $\mathcal{V}_{p,p}$, respectively.

The matrix Langevin distribution ($\mathcal{ML}$-distribution) is a widely used probability distribution on $\mathcal{V}_{n,p}$ (Mardia and Jupp, 2009; Chikuse, 2012; Lin et al., 2017). This distribution is also known as Von Mises-Fisher matrix distribution (Khatri and Mardia, 1977). As defined in Chikuse (2012), the probability density function of the matrix Langevin distribution (with respect to the normalized Haar measure μ on $\mathcal{V}_{n,p}$) parametrized by $F \in \mathbb{R}^{n\times p}$, is

$$f_{\mathcal{ML}}(X\,;\,F) = \frac{etr(F^T X)}{{}_0F_1\left(\frac{n}{2}, \frac{F^T F}{4}\right)}, \tag{2.1}$$

where $etr(\cdot) = \exp(trace(\cdot))$ and the normalizing constant, ${}_0F_1(n/2, F^T F/4)$, is the hypergeometric function of order $n/2$ with the matrix argument $F^T F/4$ (Herz, 1955;

James, 1964; Muirhead, 1975; Gupta and Richards, 1985; Gross and Richards, 1987, 1989; Butler and Wood, 2003; Koev and Edelman, 2006; Chikuse, 2012). In this article, we consider a different parametrization of the parameter matrix F in terms of its singular value decomposition (SVD). In particular, we subscribe to the specific form of unique SVD defined in Chikuse (2012) ((1.5.8) in Chikuse (2012)),

$$F = MDV^T,$$

where $M \in \widetilde{\mathcal{V}}_{n,p}$, $V \in \mathcal{V}_{p,p}$, and D is the diagonal matrix with diagonal entries $\boldsymbol{d} = (d_1, d_2, \cdots, d_p) \in \mathcal{S}_p$. Here $\widetilde{\mathcal{V}}_{n,p} = \{X \in \mathcal{V}_{n,p} : X_{1,j} \geq 0 \quad \forall\, j = 1, 2, \cdots, p\}$ and $\mathcal{S}_p = \left\{(d_1, \ldots, d_p) \in \mathbb{R}_+^p : 0 < d_p < \cdots < d_1 < \infty\right\}$. Henceforth, we shall use the phrase "unique SVD" to refer to this specific form of SVD. Khatri and Mardia (1977) (page 96) shows that the function ${}_0F_1(n/2, F^TF/4)$ depends only on the eigenvalues of the matrix F^TF, i.e.,

$${}_0F_1\left(\frac{n}{2}, \frac{F^TF}{4}\right) = {}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right).$$

As a result, we reparametrize the $\mathcal{ML}$ density as

$$f_{\mathcal{ML}}(X; (M, \boldsymbol{d}, V)) = \frac{etr(VDM^TX)}{{}_0F_1(\frac{n}{2}, \frac{D^2}{4})}\, \mathbb{I}(M \in \widetilde{\mathcal{V}}_{n,p}, \boldsymbol{d} \in \mathcal{S}_p, V \in \mathcal{V}_{p,p}).$$

This parametrization ensures identifiability of all the parameters $M, \boldsymbol{d}$ and V. With regard to interpretation, the mode of the distribution is MV^T and $\boldsymbol{d}$ represents the concentration parameter (Chikuse, 2003). For notational convenience we omit the indicator function and write the $\mathcal{ML}$ density as

$$f_{\mathcal{ML}}(X; (M, \boldsymbol{d}, V)) = \frac{etr(VDM^TX)}{{}_0F_1(\frac{n}{2}, \frac{D^2}{4})}, \tag{2.2}$$

where it is understood that $M \in \widetilde{\mathcal{V}}_{n,p}, \boldsymbol{d} \in \mathcal{S}_p, V \in \mathcal{V}_{p,p}$. The parametrization with $M, \boldsymbol{d}$ and V enables us to represent the intractable hypergeometric function of a matrix argument as a function of vector $\boldsymbol{d}$, the diagonal entries of D, paving a path for an efficient posterior inference procedure.

We note in passing that an alternative parametrization through polar decomposition with $F = MK$ (Mardia and Jupp, 2009) may pose computational challenges since the elliptical part K lies on a positive semi-definite cone and inference on positive semi-definite cone is not straightforward (Hill and Waters, 1987; Bhatia, 2009; Schwartzman, 2006).

3 Conjugate Prior for the $\mathcal{ML}$-Distribution

In the context of the exponential family of distributions, Diaconis and Ylvisaker (1979) (DY) provides a standard procedure to obtain a class of conjugate priors when the distribution is represented through its natural parametrization (Casella and Berger, 2002). Unfortunately, for the $\mathcal{ML}$ distribution, the DY theorem can not be applied directly, as demonstrated next. We therefore develop, in Section 3.2, two novel classes of priors and present a detailed investigation of their properties.

3.1 Inapplicability of DY Theorem for Construction of Priors for the $\mathcal{ML}$-Distribution

In order to present the arguments in this section, we introduce notations P_θ, $x_{_A}$, μ, and μ_A, that are directly drawn from Diaconis and Ylvisaker (1979). In brief, P_θ denotes the probability measure that is absolutely continuous with respect to an appropriate σ-finite measure μ on a convex subset of the Euclidean space, $\mathbb{R}^d$. In the case of the $\mathcal{ML}$ distribution, μ is the Haar measure defined on the Stiefel manifold. The symbol $\mathcal{X}$ denotes the interior of the support of the measure μ. As shown in Hornik and Grün (2013) $\mathcal{X} := \{X : \|X\|_2 < 1\}$ for the case of the $\mathcal{ML}$ distribution. According to the assumptions of DY $\int_{\mathcal{X}} dP_\theta(X) = 1$ (see paragraph after (2.1), page 271 in Diaconis and Ylvisaker (1979)). In the current context, P_θ is the probability measure associated with the $\mathcal{ML}$ distribution. Therefore,

$$\int_{\mathcal{X}} dP_\theta(X) = \int_{\mathcal{X}} f_{\mathcal{ML}}(X)\, \mu(dX) = 0,$$

which violates the required assumption mentioned above. Secondly, in the proof of Theorem 1 in Diaconis and Ylvisaker (1979) DY construct a probability measure restricted to a measurable set A as follows.

$$\mu_A(B) = \frac{\mu(A \cap B)}{\mu(A)}, \text{ where } \mu(A) > 0.$$

Considering the notation $x_{_A} = \int Z\ \mu_A(dZ)$ for any measurable set A, the proof of Theorem 1 in Diaconis and Ylvisaker (1979) relies on the existence of a sequence of measurable sets $\{A_j\}_{j\geq 1}$ and corresponding points $\left\{x_{A_j}\right\}_{j\geq 1}$ that are required to be dense in $supp(\mu)$, the support of the measure μ (see line after (2.4) on page 272 in Diaconis and Ylvisaker (1979)). It can be shown that a similar construction in the case of the $\mathcal{ML}$ distribution would lead to a $x_{_A}$ where $x_{_A}$ does not belong to $supp(\mu)$, the Stiefel manifold. Therefore, the mentioned set of points $\left\{x_{A_j}\right\}_{j\geq 1}$ that are dense in $supp(\mu)$ does not exist for the case of the $\mathcal{ML}$ distribution.

Together, the two observations make it evident that Theorem 1 in Diaconis and Ylvisaker (1979) is not applicable for constructing conjugate priors for the $\mathcal{ML}$ distribution. We would like to point out that the construction of the class of priors in Hornik and Grün (2013) is based on a direct application of DY, which is not entirely applicable for the $\mathcal{ML}$-distribution. On the other hand, the idea of constructing a conjugate prior on the natural parameter F followed by a transformation, involves calculations of a complicated Jacobian term (Hornik and Grün, 2013). Hence the class of priors obtained via this transformation lacks interpretation of the corresponding hyperparameters.

3.2 Two Novel Classes of Conjugate Priors

Let μ denote the normalized Haar measure on $\mathcal{V}_{n,p}$, μ_2 denote the normalized Haar measure on $\mathcal{V}_{p,p}$, and μ_1 denote the Lebesgue measure on $\mathbb{R}^p_+$. For the parameters of the $\mathcal{ML}$-distribution, we define the prior density with respect to the product measure $\mu \times \mu_1 \times \mu_2$ on the space $\mathcal{V}_{n,p} \times \mathbb{R}^p_+ \times \mathcal{V}_{p,p}$.

Definition 1. *The probability density function of the joint conjugate prior on the parameters $M, \boldsymbol{d}$ and V for the $\mathcal{ML}$ distribution is proportional to*

$$g(M, \boldsymbol{d}, V\,;\,\nu, \Psi) = \frac{etr\left(\nu\, V D M^T \Psi\right)}{\left[{}_0F_1(\frac{n}{2}, \frac{D^2}{4})\right]^{\nu}}, \tag{3.1}$$

as long as $g(M, \boldsymbol{d}, V\,;\,\nu, \Psi)$ is integrable. Here $\nu > 0$ and $\Psi \in \mathbb{R}^{n\times p}$.

Henceforth, we refer to the joint distribution corresponding to the probability density function in Definition 1 as the joint conjugate prior distribution (*JCPD*). We use the terminology, joint conjugate prior class (*JCPC*) when we use

$$(M, \boldsymbol{d}, V) \sim JCPD\,(\cdot\,;\nu, \Psi)\,, \tag{3.2}$$

as a prior distribution for the parameters of the $\mathcal{ML}$-distribution. Although, the *JCPC* has some desirable properties (see Theorem 5 and Section 5.2), it may not be adequately flexible to incorporate prior knowledge about the parameters if the strength of prior belief is not uniform across the different parameters. For example, if a practitioner has strong prior belief for the values of M but is not very certain about parameters $\boldsymbol{d}$ and V, then *JCPC* may not be the optimal choice. Also, the class of joint prior defined in Definition 1 corresponds to a dependent prior structure for the parameters M, $\boldsymbol{d}$ and V. However, it is customary to use independent prior structure for parameters of curved exponential families (Casella and Berger, 2002; Gelman et al., 2014; Khare et al., 2017). Consequently, we also develop a class of conditional conjugate prior where we assume independent priors on the parameters M, $\boldsymbol{d}$ and V. This class of priors are flexible enough to incorporate prior knowledge about the parameters even when the strength of prior belief differs across different parameters.

It is easy to see that the conditional conjugate priors for both M and V are $\mathcal{ML}$-distributions whereas the following definition is used to construct the conditional conjugate prior for $\boldsymbol{d}$.

Definition 2. *The probability density function of the conditional conjugate prior for $\boldsymbol{d}$ with respect to the Lebesgue measure on $\mathbb{R}^p_+$ is proportional to*

$$g(\boldsymbol{d}\,;\,\nu, \boldsymbol{\eta}, n) = \frac{\exp(\nu\, \boldsymbol{\eta}^T \boldsymbol{d})}{\left[{}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)\right]^{\nu}}, \tag{3.3}$$

as long as $g(\boldsymbol{d}\,;\,\nu, \boldsymbol{\eta}, n)$ is integrable. Here $\nu > 0$, $\boldsymbol{\eta} \in \mathbb{R}^p$ and $n \geq p$.

Note that $g(\boldsymbol{d}\,;\,\nu, \boldsymbol{\eta})$ is a function of n as well. However we do not vary n anywhere in our construction, and thus we omit reference to n in the notation for $g(\boldsymbol{d}\,;\,\nu, \boldsymbol{\eta})$.

Henceforth we use the terminology, conditional conjugate prior distribution for $\boldsymbol{d}$ (*CCPD*) to refer to the probability distribution corresponding to the probability density function in Definition 2. We use the phrase conditional conjugate prior class (*CCPC*),

to refer to the following structure of prior distributions

$$\begin{aligned} M &\sim \mathcal{ML}\left(\cdot;\, \xi^M, \xi^D, \xi^V\right), \\ \boldsymbol{d} &\sim CCPD\left(\cdot;\, \nu, \boldsymbol{\eta}\right), \\ V &\sim \mathcal{ML}\left(\cdot;\, \gamma^M, \gamma^D, \gamma^V\right), \end{aligned} \tag{3.4}$$

where $M, \boldsymbol{d}, V$ are assumed to be independent apriori. As per Definitions 1 and 2, the integrability of the kernels mentioned in (3) and (5) are critical to prove the propriety of the proposed class of priors. In light of this, Theorem 1 and Theorem 2 provide conditions on ν, Ψ and $\boldsymbol{\eta}$ for $g(M, \boldsymbol{d}, V\,;\, \nu, \Psi)$ and $g(\boldsymbol{d}\,;\, \nu, \boldsymbol{\eta})$ to be integrable, respectively.

Theorem 1. *Let $M \in \mathcal{V}_{n,p}$, $V \in \mathcal{V}_{p,p}$ and $\boldsymbol{d} \in \mathbb{R}^p_+$. Let $\Psi \in \mathbb{R}^{n\times p}$ with $n \geq p$, then for any $\nu > 0$,*

(a) If $\|\Psi\|_2 < 1$, then

$$\int_{\mathcal{V}_{n,p}} \int_{\mathcal{V}_{p,p}} \int_{\mathbb{R}^p_+} g(M, \boldsymbol{d}, V\,;\, \nu, \Psi)\; d\mu_1(\boldsymbol{d})\; d\mu_2(V)\; d\mu(M) < \infty,$$

(b) If $\|\Psi\|_2 > 1$, then

$$\int_{\mathcal{V}_{n,p}} \int_{\mathcal{V}_{p,p}} \int_{\mathbb{R}^p_+} g(M, \boldsymbol{d}, V\,;\, \nu, \Psi)\; d\mu_1(\boldsymbol{d})\; d\mu_2(V)\; d\mu(M) = \infty,$$

where $g(M, \boldsymbol{d}, V; \nu, \Psi)$ is defined in Definition 1.

The conditions mentioned in this theorem do not span all cases; we have not addressed the case where $\|\Psi\|_2 = 1$. As far as statistical inference for practical applications is concerned, we may not have to deal with the case where $\|\Psi\|_2 = 1$ as the hyper-parameter selection procedure (see Section 4) and posterior inference (even in the case of uniform improper prior, see Section 5.3) only involve cases with $\|\Psi\|_2 < 1$. We therefore postpone further investigation into this case as a future research topic of theoretical interest.

Theorem 2. *Let $\boldsymbol{d} \in \mathbb{R}^p_+$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p) \in \mathbb{R}^p$ and n be any integer with $n \geq p$. Then for any $\nu > 0$,*

$$\int_{\mathbb{R}^p_+} g(\boldsymbol{d}; \nu, \boldsymbol{\eta}, n)\; d\mu_1(\boldsymbol{d}) < \infty,$$

if and only if $\max_{1\leq j\leq p} \eta_j < 1$, where $g(\boldsymbol{d}; \nu, \boldsymbol{\eta}, n)$ is as defined in Definition 2.

We can alternatively parametrize the *CCPD* class of densities by the following specification of the probability density function,

$$f(\boldsymbol{d}\,;\, \nu, \boldsymbol{\eta}) \propto \frac{\exp\left(\sum_{j=1}^p \eta_j d_j\right)}{\left[{}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)\right]^\nu},$$

where $\max_{1\leq j\leq p}\eta_j < \nu$. In this parametrization, if we consider the parameter choices, $\nu = 0$ and $\boldsymbol{\beta} := -\boldsymbol{\eta}$, then the resulting probability distribution corresponds to the *Exponential* distribution with rate parameter $\boldsymbol{\beta}$.

Properties of the *CCPD* and *JCPD* Class of Distributions It is important to explore the properties for the *CCPD* and *JCPD* class of distributions in order to use them in an effective manner. Intuitive interpretations of the parameters $\nu, \boldsymbol{\eta}, \Psi$ are desirable, for example, for hyper-parameter selection. Due to conjugacy, Bayesian analysis will lead to posterior distributions involving *JCPD* and *CCPD*, and therefore, it is necessary to identify features that are required to develop practicable computation schemes for posterior inference. The following four theorems establish some crucial properties of the *CCPD* and *JCPD* class of distributions.

Theorem 3. *Let* $\boldsymbol{d} \sim CCPD(\cdot; \nu, \boldsymbol{\eta})$ *for* $\nu > 0$ *and* $\max_{1\leq j\leq p}\eta_j < 1$ *where* $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$. *Then*

(a) The distribution of $\boldsymbol{d}$ *is log-concave.*

(b) The distribution of $\boldsymbol{d}$ *has a unique mode if* $\eta_j > 0$ *for all* $j = 1, 2, \cdots, p$. *The mode of the distribution is given by* $\mathbf{m}_{\boldsymbol{\eta}} = \mathbf{h}^{-1}(\boldsymbol{\eta})$, *where the function* $\mathbf{h}(\boldsymbol{d})$ *is defined as follows,* $\mathbf{h}(\boldsymbol{d}) := (h_1(\boldsymbol{d}), h_2(\boldsymbol{d}), \cdots, h_p(\boldsymbol{d}))^T$ *with*

$$h_j(\boldsymbol{d}) := \left(\frac{\partial}{\partial d_j}\,{}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)\right) / {}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right).$$

Notably, the mode of the distribution is characterized by the parameter $\boldsymbol{\eta}$ and does not depend on the parameter ν. The proof of the theorem relies on a few nontrivial properties of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$, i.e., the hyper-geometric function of a matrix argument, that we have established in the supplementary material Section 1. It is easy to see that the function $\mathbf{h}^{-1}$ is well defined as the function $\mathbf{h}$ is strictly increasing in all its coordinates. Even though subsequent theoretical developments are based on the formal definition and theoretical properties of $\mathbf{h}^{-1}$ and $\mathbf{h}$ functions, numerical computation of the functions are tricky. The evaluation of the functions depend on reliable computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ and all its partial derivatives. In Section 6.2, we provide a reliable and theoretically sound computation scheme for these functions.

On a related note, it is well known that log-concave densities correspond to unimodal distributions if the sample space is the entire Euclidean space (Ibragimov, 1956; Dharmadhikari and Joag-Dev, 1988; Doss and Wellner, 2016). However, the mode of the distribution may not necessarily be at a single point. part (b) of Theorem 3 asserts that the *CCPD* has a single point mode. Moreover, the sample space of *CCPD* is $\boldsymbol{d} \in \mathbb{R}_+^p$, which merely encompasses the positive quadrant and not the whole of the p dimensional Euclidean space. Hence general theories developed for $\mathbb{R}^p$ (or $\mathbb{R}$) do not apply. In fact, when $\eta_j \leq 0$, the density defined in Definition 2 is decreasing as a function of d_j on the set $\mathbb{R}_+$ and the mode does not exist as $\mathbb{R}_+$ does not contain the point 0. In all, part (b) of Theorem 3 does not immediately follow from part (a) and requires additional effort to demonstrate.

In order to introduce the notion of "concentration" for the *CCPD* class of distributions we require the concept of a level set. Let the unnormalized probability density function for the *CCPD* class of distributions, $g(\boldsymbol{x};\nu,\boldsymbol{\eta})$ (see Definition 5), achieve its maximum value at $\mathbf{m}_{\boldsymbol{\eta}}$ (part(b) of Theorem 3 ensures that $\mathbf{m}_{\boldsymbol{\eta}}$ is a unique point) and let

$$\mathcal{S}_l = \left\{\boldsymbol{x} \in \mathbb{R}_+^p : g(\boldsymbol{x};1,\boldsymbol{\eta})/g(\mathbf{m}_{\boldsymbol{\eta}};1,\boldsymbol{\eta}) > l\right\} \tag{3.5}$$

be the level set of level l containing the mode $\mathbf{m}_{\boldsymbol{\eta}}$ where $0 \leq l < 1$. To define the level set we could have used $g(\boldsymbol{x};\nu_0,\boldsymbol{\eta})$ for any fixed value of $\nu_0 > 0$ instead of $g(\boldsymbol{x};1,\boldsymbol{\eta})$. However, without loss of generality, we choose $\nu_0 = 1$.

Let $P_\nu(\cdot;\boldsymbol{\eta})$ denote the probability distribution function corresponding to the $CCPD(\cdot;\nu,\boldsymbol{\eta})$ distribution. According to Theorem 3, for a fixed $\boldsymbol{\eta} \in \mathbb{R}^p$, all distributions in the class $\{P_\nu(\cdot;\boldsymbol{\eta}) : \nu > 0\}$ have the mode located at the point $\mathbf{m}_{\boldsymbol{\eta}}$.

Theorem 4. *Let $\boldsymbol{d}_\nu \sim CCPD(\cdot;\nu,\boldsymbol{\eta})$ for a fixed $\boldsymbol{\eta} \in \mathbb{R}^p$ with $\mathbf{m}_{\boldsymbol{\eta}}$ being the mode of the distribution. If $P_\nu(\cdot;\boldsymbol{\eta})$ denotes the probability distribution function corresponding to $\boldsymbol{d}_\nu$, then*

(a) $P_\nu(\mathcal{S}_l;\boldsymbol{\eta})$ is an increasing function of ν for any level set $\mathcal{S}_l$ with $l \in (0,1)$,

(b) For any open set $\mathcal{S} \subset \mathbb{R}_+^p$ containing $\mathbf{m}_{\boldsymbol{\eta}}$, $P_\nu(\boldsymbol{d} \in \mathcal{S};\boldsymbol{\eta})$ goes to 1 as $\nu \to \infty$.

The major impediment to proving Theorem 4 arises from the intractability of the normalizing constant of the $CCPD(\cdot;\nu,\boldsymbol{\eta})$ distribution. Although involved, the proof essentially uses the log convexity of ${}_0F_1\left(\frac{n}{2},\frac{D^2}{4}\right)$ to get around this intractability.

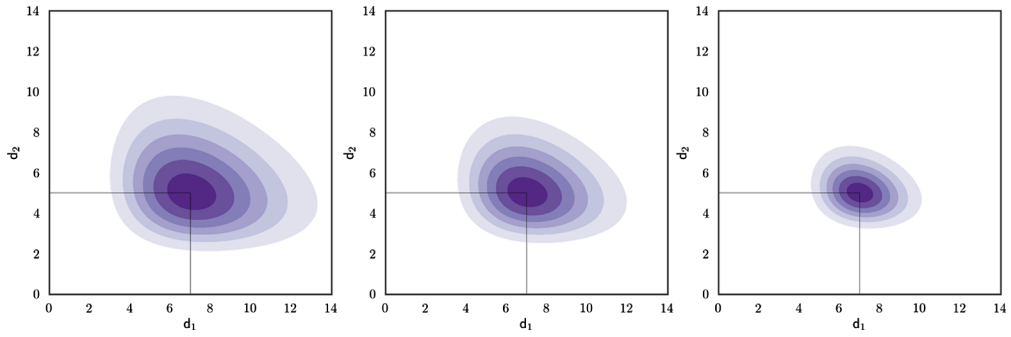
From Theorem 4, it is clear that the parameter ν relates to the concentration of the probability around the mode of the distribution. Larger values of ν imply larger concentration of probability near the mode of the distribution.

Definition 3. *In the context of the probability distribution $CCPD(\cdot\,;\,\boldsymbol{\eta},\nu)$, the parameters $\boldsymbol{\eta}$ and ν are labeled as the "modal parameter" and the "concentration parameter", respectively.*

In Figure 1, we display three contour plots of the $CCPD(\cdot\,;\nu,\boldsymbol{\eta})$ distribution with $\boldsymbol{\eta} = (0.85, 0.88)$. Note that the corresponding mode of the distribution is $\mathbf{h}^{-1}(0.85, 0.88) = (7,5)$ for all three plots. We can observe the implication of part (b) of Theorem 3 as the "center" of the distributions are the same. Contrastingly, it can be observed that the "spread" of the distributions decrease as the value of the parameter ν increases, as implied by Theorem 4.

Theorem 5. *Let $(M,\boldsymbol{d},V) \sim JCPD(\cdot;\nu,\Psi)$ for some $\nu > 0$ and $\|\Psi\|_2 < 1$. If $\Psi = M_\Psi D_\Psi V_\Psi^T$ is the unique SVD of Ψ with $\boldsymbol{d}_\Psi$ being the diagonal elements of D_Ψ, then the unique mode of the distribution is given by $(M_\Psi, \mathbf{h}^{-1}(\boldsymbol{d}_\Psi), V_\Psi)$ where the function $\boldsymbol{d} \to \mathbf{h}(\boldsymbol{d})$ is as defined in Theorem 3.*

Note that the mode of the distribution is characterized by the parameter Ψ and does not depend on the parameter ν. The proof of the theorem depends crucially on a strong result, a type of rearrangement inequality proved in Kristof (1969).



(a) $\nu = 10$ (b) $\nu = 20$ (c) $\nu = 35$

Figure 1: Density plots of $CCPD(\cdot\,;\,\nu, \boldsymbol{\eta})$ for different values of ν where $\boldsymbol{\eta} = (0.89, 0.85)$. Mode of the distributions are located at the point $(7, 5)$.

For the concentration characterization of *JCPD*, we define the level sets in the context of the *JCPD* distribution. Let the unnormalized probability density function for the *JCPD* class of distributions, $g(M, \boldsymbol{d}, V; \nu, \Psi)$, achieve its maximum value at the point $(\hat{M}, \hat{\boldsymbol{d}}, \hat{V})$ (see Theorem 5) and

$$\mathcal{A}_l = \left\{ (M, \boldsymbol{d}, V) \in \mathcal{V}_{n,p} \times \mathbb{R}^p_+ \times \mathcal{V}_{p,p} : g(M, \boldsymbol{d}, V; 1, \Psi)/g(\hat{M}, \hat{\boldsymbol{d}}, \hat{V}; 1, \Psi) > l \right\}$$

be the level set of level l from some $l \in (0, 1)$. The following theorem characterizes the concentration property of the *JCPD* distribution.

Theorem 6. *Let $(M, \boldsymbol{d}, V) \sim JCPD(\cdot; \nu, \Psi)$, where $\|\Psi\|_2 < 1$. If $P_\nu(\cdot\,;\,\Psi)$ denotes the probability distribution function corresponding to the distribution $JCPD(\cdot; \nu, \Psi)$, then*

(a) $P_\nu(\mathcal{A}_l\,;\,\Psi)$ is a strictly increasing function of ν for any level set $\mathcal{A}_l$ with $l \in (0, 1)$.

(b) For any open set $\mathcal{A} \subset \mathcal{V}_{n,p} \times \mathbb{R}^p_+ \times \mathcal{V}_{p,p}$ containing the mode of the distribution, $P_\nu(\mathcal{A}\,;\,\Psi)$ tends to 1 as $\nu \to \infty$.

(c) The conditional distribution of M given $(\boldsymbol{d}, V)$ and V given $(M, \boldsymbol{d})$ are $\mathcal{ML}$ distributions whereas the conditional distribution of $\boldsymbol{d}$ given (M, V) is a CCPD distribution.

Parts (a) and (b) of the above theorem characterize the concentration whereas part (c) relates *CCPD* to the *JCPD* class of distributions. Part(c) also motivates the development of a sampling procedure for the *JCPD* distribution. The proof of part (a) Theorem 6 is similar to that of the proof of Theorem 4. The proof for part (b) of Theorem 6 is more involved and depends on several key results, including the rearrangement inequality by Kristof (1969), the log convexity of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$, and the fact that $g(\mathbf{h}^{-1}(\boldsymbol{\eta})\,;\,\nu, \boldsymbol{\eta})$, the value of the unnormalized *CCPD* density at the mode, is a strictly increasing function of the parameter $\boldsymbol{\eta}$.

Note that unlike in the case of the *CCPD* distribution, we do not attempt to establish the log concavity of *JCPD*, the reason being that the underlying probability space $\mathcal{V}_{n,p} \times \mathbb{R}_+^p \times \mathcal{V}_{p,p}$ is non-convex. Nevertheless, it is evident that beyond a certain distance (based on a suitable metric on $\mathcal{V}_{n,p} \times \mathbb{R}_+^p \times \mathcal{V}_{p,p}$) the value of the density drops monotonically as one moves farther away from the center. Based on the characteristics of the parameters ν and Ψ of the *JCPD* class of distributions, we have the following definitions.

Definition 4. *The parameters Ψ and ν in the distribution JCPD are labeled the "modal" parameter and the "concentration" parameter, respectively.*

Interestingly, both distributions *CCPD* and *JCPD* are parameterized by two parameters, one controlling the center and the other characterizing the probability concentration around that center. One may therefore visualize the distributions in a fashion similar to that of the multivariate Normal distribution controlled by the mean and variance parameters. This intuitive understanding can help practitioners select hyper-parameter values when conducting a Bayesian analysis with the *CCPD* and *JCPD* distributions.

Thus far we have established properties of *CCPD* and *JCPD* that relate to basic features of these distributions. Additional properties, which are required for a Markov chain monte carlo (MCMC) sampling scheme, are developed in Section 5.1.

4 Hyperparameter Selection Procedure

4.1 Informative Prior

We now present procedures for the selection of hyperparameter values aimed at incorporating prior beliefs about the parameters $(M, \boldsymbol{d}, V)$. Consider the scenario where a practitioner has the prior belief that the values for the parameters $M, \boldsymbol{d}, V$ are close to $M_{belief}, \boldsymbol{d}_{belief}, V_{belief}$, respectively. A standard approach to incorporating this prior knowledge is to select the hyper-parameter values in such a manner that the mode of the corresponding prior distribution becomes $M_{belief}, \boldsymbol{d}_{belief}, V_{belief}$. In order to achieve this in the current context, we first compute $\tilde{\boldsymbol{\eta}} = h(\boldsymbol{d}_{belief})$ where $h(\cdot)$ is defined in (2.8) in the supplementary material. Note that we always get a feasible $\tilde{\boldsymbol{\eta}}$ for every real $\boldsymbol{d}_{belief} \in \mathcal{S}_p$.

In the case of the *CCPC* class of priors, we choose $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$, $\xi^M = M_{belief}$, $\gamma^M = V_{belief}$, $\xi^V = \mathbf{I}_p, \gamma^V = \mathbf{I}_p$ in (3.4). Theorem 3 guarantees that the above hyper-parameter specifications yields a prior distribution that has mode at $(M_{belief}, \boldsymbol{d}_{belief}, V_{belief})$. From Theorem 3, we also see that larger values of the hyper-parameter ν lead to larger concentration of the prior probability around the mode. The hyper-parameters ξ^D and γ^D play a similar role for the $\mathcal{ML}$ distribution. Hence the hyper parameters ν, ξ^D and γ^D are chosen to have larger values in case the practitioner has a higher confidence in the prior belief.

In the case of the *JCPC* class of priors, we apply Theorem 5 to construct *JCPD* (see (3.2)) with mode at $M_{belief}, \boldsymbol{d}_{belief}, V_{belief}$. In particular, we set $\Psi = M_{belief} D_{\tilde{\boldsymbol{\eta}}} (V_{belief})^T$ where $D_{\tilde{\boldsymbol{\eta}}}$ is the diagonal matrix with diagonal elements $\tilde{\boldsymbol{\eta}} = h(\boldsymbol{d}_{belief})$. Using the

concentration characterization described in Theorem 5, the practitioner may choose the value of the hyper-parameter ν appropriately, where a larger value for the parameter ν implies greater confidence in the prior belief.

It is noteworthy that for both the *JCPC* and *CCPC* class of priors, there is an intimate connection between the sample size and the interpretation of the hyper-parameter ν. As a heuristic one may envisage ν as incorporating "information" equivalent to ν many historic observations of the model.

4.2 Uniform Improper Prior

In the case where the practitioner does not have a prior belief about the parameter values, an automatic procedure for hyper-parameter selection can be helpful. In this and the next subsection, we discuss two automatic procedures to select the values of the hyper-parameters. In the absence of prior information, usage of uniform prior is common in the literature. In the context of the current model, for the *JCPC* and *CCPC* class of distributions, the prior for the parameters $(M, \boldsymbol{d}, V)$, is called a uniform prior if

$$g(M, \boldsymbol{d}, V; \nu, \Psi) \propto 1 \text{ and}$$
$$f_{\mathcal{ML}}(M\,; \xi^M, \xi^D, \xi^V) g(\boldsymbol{d}; \nu, \boldsymbol{\eta}) f_{\mathcal{ML}}(V\,;\, \gamma^M, \gamma^D, \gamma^V) \propto 1.$$

Both classes of priors *JCPC* and *CCPC* are flexible enough to accommodate a uniform prior. For *JCPC*, this can be achieved by setting $\nu = 0$ in (3.2). Correspondingly, for the *CCPC* class, the uniform prior can be constructed by choosing $\nu = 0$, $\xi^D = \mathbf{0}$ and $\gamma^D = \mathbf{0}$ in (3.4). Note that the resulting uniform prior is improper in nature as the above choices of hyper parameters do not lead to a proper probability distribution. Hence, it is necessary to check the propriety of the resulting posterior (see Section 5.3 for more details).

4.3 Empirical Prior

Another widely used automatic method is to use empirical information contained in the data to select appropriate values of the hyper-parameters. Let $W_1, W_2, \ldots W_N$ be independent and identically distributed samples drawn from $\mathcal{ML}(\cdot\,; M, \boldsymbol{d}, V)$. Consider the sample mean, $\overline{W} = (\sum_{i=1}^{N} W_i)/N$. Let the unique SVD of the sample mean be $\overline{W} = M_{\overline{W}} D_{\overline{W}} V_{\overline{W}}$. Construct candidate values $M_{belief} = M_{\overline{W}}$, $V_{belief} = V_{\overline{W}}$ and $\tilde{\boldsymbol{\eta}}$ as the diagonal elements of $D_{\overline{W}}$. One can set $\Psi = \overline{W}$ as the hyper-parameter in the case of the *JCPC* prior. In the case of the *CCPC* class of priors, one can choose $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$, and for the hyper-parameters related to M and V, apply the same procedure as discussed previously in this section. For both classes of priors, a value for ν that is less than or equal to 10 percent of the sample size N, is recommended.

Example 1. *Let the practitioner have the following prior belief for the values of the parameters* $M, \boldsymbol{d}, V$,

$$M_{belief} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \; \boldsymbol{d}_{belief} = \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \; V_{belief} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

As described previously in this section, we can compute $\tilde{\boldsymbol{\eta}} = \mathbf{h}(7,5) = (0.89, 0.85)$. *Hence, for the JCPC class of priors, we choose the hyper-parameter values*

$$\tilde{\Psi} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.89 & 0 \\ 0 & 0.85 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^T = \begin{bmatrix} 0.89 & 0 \\ 0 & 0.85 \\ 0 & 0 \end{bmatrix},$$

to ensure that $JCPD(\cdot\,; \tilde{\Psi}, \nu)$ *has mode at* $M_{belief}, \boldsymbol{d}_{belief}, V_{belief}$ *for all values of* $\nu > 0$. *The value of the hyper-parameter* ν *should be chosen according to the strength of the prior belief. In Figure 1, we display the resulting conditional distribution for* $\boldsymbol{d}$ *given* M, V. *Figure 1 shows that the "center" of the distribution is located at* $(7,5)$. *Figure 1 also displays the "spread" of the distribution around the mode when using* $\nu = 10$, $\nu = 20$ *and* $\nu = 35$.

5 Properties of Posterior

The derivation of the posterior distributions for the *JCPC* and *CCPC* class of priors is straightforward since they were built with conjugacy in mind, which then entails that the posterior distributions lie in the corresponding classes. However, inference for the resulting posterior distributions is challenging because not only are the normalizing constants intractable for both the *JCPD* and *CCPD* distributions, but also, the unnormalized version of the corresponding density functions involve ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$. We first focus our attention on developing properties of the posterior distribution when involving *JCPC* and *CCPC* priors. In particular, we derive explicit forms of the posterior conditionals under different prior settings, the linearity of the posterior mode parameters and the strong consistency of the posterior mode.

5.1 Posterior Conditionals

Let $W_1, W_2, \ldots W_N$ be independent and identically distributed samples drawn from $\mathcal{ML}(\cdot\,; M, \boldsymbol{d}, V)$. Let $\overline{W} = \sum_{i=1}^{N} W_i/N$. The likelihood of the data is

$$\prod_{i=1}^{N} \frac{etr(VDM^T W_i)}{{}_0F_1(\frac{n}{2}, \frac{D^2}{4})}. \tag{5.1}$$

First, let us assume a *JCPD* prior with parameters ν and Ψ. Theorem 5 not only implies that the posterior has a unique mode, but also provides an expression for the mode. Furthermore, we see that the corresponding posterior distribution is *JCPD* with concentration $(\nu + N)$ and posterior modal parameter $\widehat{\Psi}_N = \left(\frac{\nu}{\nu+N}\Psi + \frac{N}{\nu+N}\overline{W}\right)$. Let $\hat{\boldsymbol{\eta}}_{\Psi_N}$ be the diagonal elements of the diagonal matrix $\hat{D}_{\Psi_N}$, where $\widehat{\Psi}_N = \hat{M}_N \hat{D}_{\Psi_N} \hat{V}_N$ is the unique SVD for $\widehat{\Psi}_N$. From Theorem 6, it follows that the full posterior conditionals for the parameters $M, \boldsymbol{d}, V$ are $\mathcal{ML}$, *CCPD* and $\mathcal{ML}$ distributions, respectively.

In Section 6 we shall use these results to construct a Gibbs algorithm. A part of the Gibbs scheme would require sampling from the relevant *CCPD* distribution, which we

propose to implement by simulating from the full conditional distribution of each of the components of $\boldsymbol{d}$ given the rest, when $\boldsymbol{d} \sim CCPD(\cdot\,; \nu, \boldsymbol{\eta})$. To refer to this conditional distribution in subsequent text, we have the following definition.

Definition 5. *Let $\nu > 0$, $\boldsymbol{\varpi} \in \mathbb{R}_+^{p-1}$ and $\boldsymbol{\eta} \in \mathbb{R}_+^p$ with $\max_{1\le j\le p} \eta_j < 1$. A random variable is defined to be distributed as $CCPD_j^\star(\cdot\,; \boldsymbol{\varpi}, \nu, \boldsymbol{\eta})$, if the corresponding probability density function (with respect to the Lebesgue measure on $\mathbb{R}$) is proportional to*

$$g_j(x;\, \boldsymbol{\varpi}, \nu, \boldsymbol{\eta}) = \frac{\exp(\nu\, \eta_j x)}{\left[{}_0F_1\left(\frac{n}{2}, \frac{(\Delta(x))^2}{4}\right)\right]^\nu},$$

where $\Delta(x)$ is a diagonal matrix with diagonal elements $(x, \boldsymbol{\varpi}) \in \mathbb{R}_+^p$.

Let $\boldsymbol{d} = (d_1, \ldots, d_p)$ be a random vector with $\boldsymbol{d} \sim CCPD(\cdot\,; \nu, \boldsymbol{\eta})$ for some $\max_{1\le j\le p} \eta_j < 1, \nu > 0$. Let $\boldsymbol{d}^{(-j)}$ be the vector containing all but the j-th component of the vector $\boldsymbol{d}$. Then the conditional distribution of d_j given $\boldsymbol{d}^{(-j)}$ is $CCPD_j^\star(\cdot\,; \boldsymbol{d}^{(-j)}, \nu, \boldsymbol{\eta})$, i.e.,

$$d_j \mid \boldsymbol{d}^{(-j)} \sim CCPD_j^\star(\cdot\,; \boldsymbol{d}^{(-j)}, \nu, \boldsymbol{\eta}).$$

Now, since the conditional posterior of $\boldsymbol{d}$ was shown to be $CCPD$, the conditional posterior distribution of $d_j \mid \boldsymbol{d}^{(-j)}, M, V, \{W_i\}_{i=1}^N$ follows a $CCPD_j^\star$ distribution.

In the case of a Bayesian analysis with a $CCPC$ prior, (3.4) and (5.1) determine the corresponding posterior distribution to be proportional to

$$\frac{etr\left(\left(V\, D\, M^T\right)\, N\, \overline{W} + G^0\, M + H^0\, V\right)}{{}_0F_1(\frac{n}{2}; D^2/4)^{\nu+N}} \exp(\nu\, \boldsymbol{\eta}^T \boldsymbol{d}), \tag{5.2}$$

where $G^0 = \xi^V\, \xi^D\, (\xi^M)^T$ and $H^0 = \gamma^V\, \gamma^D\, (\gamma^M)^T$. The conditional probability density for the posterior distribution of $\boldsymbol{d}$ given M, V, $\{W_i\}_{i=1}^N$ is proportional to

$$\frac{\exp\left((\nu+N)\left(\frac{\nu}{\nu+N}\boldsymbol{\eta} + \frac{N}{\nu+N}\boldsymbol{\eta}_{\overline{W}}\right)^T \boldsymbol{d}\right)}{\left[{}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)\right]^{\nu+N}}, \tag{5.3}$$

where $\boldsymbol{\eta}_{\overline{W}} = (Y_{1,1}, \cdots, Y_{p,p})$ with $Y = M^T \overline{W} V$. It follows that the conditional posterior distribution of $\boldsymbol{d}$ given $M, V, \{W_i\}_{i=1}^N$ is $CCPD(\cdot\ ;\ \hat{\nu}_N, \hat{\eta}_N)$ where $\hat{\nu}_N = \nu + N$ and $\hat{\eta}_N = \left(\frac{\nu}{\nu+N}\boldsymbol{\eta} + \frac{N}{\nu+N}\boldsymbol{\eta}_{\overline{W}}\right)$. The conditional posterior distributions $M \mid \boldsymbol{d}, V, \{W_i\}_{i=1}^N$ and $V \mid \boldsymbol{d}, M, \{W_i\}_{i=1}^N$ are $\mathcal{ML}$ distributions.

5.2 Linearity of Posterior Modal Parameter

We observe that the posterior modal parameter is a convex combination of the prior modal parameter and the sample mean when applying the $JCPC$ class of priors. In particular, from Section 5.1 we get

$$\hat{\Psi}_N = \left(\frac{\nu}{\nu+N}\Psi + \frac{N}{\nu+N}\overline{W}\right).$$

In a similar fashion, we observe from (5.3) that the modal parameter for the conditional posterior distribution of $\boldsymbol{d}$ given $M, V, \{W_i\}_{i=1}^N$ is a convex combination of the prior modal parameter and an appropriate statistic of the sample mean. We should point out here that the posterior linearity of the *natural parameter* of an exponential family distribution directly follows from Diaconis and Ylvisaker (1979). However, in our parametrization, the ML density is a curved exponential family of its parameters and posterior linearity appears to hold for the "modal parameter".

5.3 Posterior Propriety When Using Uniform Improper Prior

In the case where a uniform improper prior is used, the corresponding posterior is proportional to

$$\frac{etr\left(N\ VDM^T\overline{W}\right)}{\left[{}_0F_1(\frac{n}{2}, \frac{D^2}{4})\right]^N}, \tag{5.4}$$

where $\overline{W} = \frac{1}{N}\sum_{i=1}^N W_i$ (see (5.1)). It follows from Theorem 1 that the function in (5.4) leads to a proper distribution, $JCPD(\cdot\ ;\ N, \overline{W})$, if $\left\|\overline{W}\right\|_2 < 1$. The following theorem outlines the conditions under which $\left\|\overline{W}\right\|_2 < 1$.

Theorem 7. *Let $W_1, \ldots, W_N$ be independent and identically distributed samples from an $\mathcal{ML}$-distribution on the space $\mathcal{V}_{n,p}$. If*

(a) $N \geq 2$, $p < n$,

(b) $N \geq 3$, $p = n \geq 3$,

then $\left\|\overline{W}\right\|_2 < 1$ with probability 1, where $\overline{W} = \frac{1}{N}\sum_{i=1}^N W_i$.

5.4 Strong Consistency of the Posterior Mode

In the case where we use a $JCPD(\cdot; \nu, \Psi)$ prior for Bayesian analysis of the data $\{W_i\}_{i=1}^N$, the corresponding posterior distribution is a $JCPD$ with concentration $\nu + N$ and posterior modal parameter $\widehat{\Psi}_N = \left(\frac{\nu}{\nu+N}\Psi + \frac{N}{\nu+N}\overline{W}\right)$ (see Section 5.1). Let $\widehat{\Psi}_N = M_\Psi D_\Psi V_\Psi^T$ be the unique SVD of $\widehat{\Psi}_N$ with $\boldsymbol{d}_\Psi$ being the diagonal elements of D_Ψ. Then from Theorem 5, the unique mode of the distribution is given by $(\hat{M}_N, \hat{\boldsymbol{d}}_N, \hat{V}_N)$ where

$$\hat{M}_N = M_\Psi, \hat{\boldsymbol{d}}_N = \mathbf{h}^{-1}(\boldsymbol{d}_\Psi) \text{ and } \hat{V}_N = V_\Psi.$$

The form of the function $\mathbf{h}(\boldsymbol{d})$ is provided in Theorem 3. The nontrivial aspect of finding the posterior mode is the computation of the function $\mathbf{h}^{-1}(\boldsymbol{d}_\Psi)$. In our applications, we use a Newton-Raphson procedure to obtain $\mathbf{h}^{-1}(\boldsymbol{d}_\Psi)$ numerically. We use large and small argument approximations for ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ (see Jupp and Mardia (1979)) to initialize the Newton-Raphson algorithm for faster convergence. Note that the success of the Newton-Raphson procedure here depends on the efficient computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ and its

partial derivatives. In Section 6.2, we provide a method to compute these functions reliably.

The following theorem demonstrates that the mode of the posterior distribution is a strongly consistent estimator for the parameters $M, \boldsymbol{d}, V$.

Theorem 8. *Let $W_1, \ldots, W_N$ be independent and identically distributed samples from $\mathcal{ML}(\cdot\ ;\ M, \boldsymbol{d}, V)$. Let $\hat{M}_N, \hat{\boldsymbol{d}}_N$ and $\hat{V}_N$ be the posterior mode when a JCPC prior is used. The statistic $\hat{M}_N, \hat{D}_N$ and $\hat{V}_N$ are consistent estimators for the parameters M, D and V. Moreover*

$$(\hat{M}_N, \hat{\boldsymbol{d}}_N, \hat{V}_N) \xrightarrow{a.s.} (M, \boldsymbol{d}, V) \text{ as } N \longrightarrow \infty,$$

where a.s. stands for almost sure convergence.

6 MCMC Sampling from the Posterior

Apart from finding the posterior mode, a wide range of statistical inference procedures including point estimation, interval estimation (see Section 8) and statistical decision making (see Section 8) can be performed with the help of samples from the posterior distribution. For the *JCPD* and *CCPD* classes of distributions, neither is it possible to find the posterior mean estimate via integration, nor can we directly generate i.i.d. samples from the distributions. We therefore develop procedures to generate MCMC samples using a Gibbs sampling procedure, which requires the results on posterior conditionals stated in Section 5.1.

It follows from Theorem 6 and Section 5.1 that under *JCPD* prior the conditional distribution of M given $\boldsymbol{d}, V$ and the conditional distribution of V given $M, \boldsymbol{d}$ are $\mathcal{ML}$ distributions, while the conditional distribution of $\boldsymbol{d}$ given M, V is *CCPD*. Consequently, the conditional distribution of $d_j \mid \boldsymbol{d}^{(-j)}, M, V, \{W_i\}_{i=1}^N$ follows a $CCPD_j^\star$ distribution (see Definition 5). Also, let us assume that the unique SVD for $\hat{\nu}_N(\hat{\Psi}_N V D) = M_{\hat{\Psi}}^M D_{\hat{\Psi}}^M (V_{\hat{\Psi}}^M)^T$ and for $\hat{\nu}_N(\hat{\Psi}_N^T M D) = M_{\hat{\Psi}}^V D_{\hat{\Psi}}^V (V_{\hat{\Psi}}^V)^T$. Also, let us denote the vector containing the diagonal element of the matrix $M^T \hat{\Psi}_N V$ to be $\boldsymbol{\eta}_{\hat{\Psi}}$. Based on the above discussion, we can now describe the algorithm as follows.

Algorithm 1 Gibbs sampling algorithm to sample from posterior when using *JCPC* prior.

1: Sample $M \mid \boldsymbol{d}, V, \{W_i\}_{i=1}^N \sim \mathcal{ML}\left(\cdot\ ;\ M_{\hat{\Psi}}^M, \boldsymbol{d}_{\hat{\Psi}}^M, V_{\hat{\Psi}}^M\right)$,

2: Sample $d_j \mid \boldsymbol{d}^{(-j)} M, V, \{W_i\}_{i=1}^N \sim CCPD_j^\star\left(\cdot\ ; \boldsymbol{d}^{(-j)}, \hat{\nu}_N, \boldsymbol{\eta}_{\hat{\Psi}}\ \right)$ for $j = 1 \ldots p$,

3: Sample $V \mid \boldsymbol{d}, V, \{W_i\}_{i=1}^N \sim \mathcal{ML}\left(\cdot\ ;\ M_{\hat{\Psi}}^V, \boldsymbol{d}_{\hat{\Psi}}^V, V_{\hat{\Psi}}^V\right)$.

If instead we use a *CCPC* prior, (see (3.4)) for Bayesian analysis of the data, then the full conditional distribution of $M, \boldsymbol{d}, V$ are $\mathcal{ML}$, *CCPD* and $\mathcal{ML}$ distributions, respectively. The steps involved in the Gibbs sampling Markov chain are then as follows.

Algorithm 2 Gibbs sampling algorithm to sample from posterior when using *CCPC* prior.

1: Sample $M \mid \boldsymbol{d}, V, \{W_i\}_{i=1}^N \sim \mathcal{ML}\left(\cdot\,;\, S_G^M, S_G^D, S_G^V\right)$,
2: Sample $d_j \mid \boldsymbol{d}^{(-j)}, M, V, \{W_i\}_{i=1}^N \sim CCPD_j^\star\left(\cdot\,;\boldsymbol{d}^{(-j)}, \hat{\nu}_N, \hat{\eta}_N,\ \right)$ for $j = 1, \ldots p$,
3: Sample $V \mid M, \boldsymbol{d}, \{W_i\}_{i=1}^N \sim \mathcal{ML}\left(\cdot\,;\, S_H^M, S_H^D, S_H^V\right)$,

where $\hat{\nu}_N$, $\hat{\eta}_N$ are defined in (5.3) and (S_G^M, S_G^D, S_G^V), (S_H^M, S_H^D, S_H^V) are the unique SVD of the matrices $(DV^T N\overline{W}^T + G^0)$ and $(DV^T N\overline{W}^T + H^0)$, respectively.

To implement the above algorithms we need to sample from the $\mathcal{ML}$ and *CCPD* distributions. For the former, we use the procedure developed in Hoff (2009) to sample from the $\mathcal{ML}$ distributions. Sampling from $CCPD_j^\star$ is much more involved and is explained in detail in the next subsection. The following result provides some theoretical guarantees that shall be useful for this specific sampler.

Theorem 9. *Let $\boldsymbol{d} \sim CCPD(\cdot; \nu, \boldsymbol{\eta})$ for some $\nu > 0$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$ where $\max_{1\le j\le p} \eta_j < 1$. Let $g_1(\cdot\,; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})$ denote the unnormalized density corresponding to $CCPD_1^\star(\cdot\,; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})$, the conditional distribution of d_1 given $(d_2, \ldots, d_p)$.*

(a) The probability density function corresponding to $CCPD_1^\star(\cdot\ ; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})$ is log-concave on the support $\mathbb{R}_+$.

(b) If $0 < \eta_1 < 1$, the distribution $CCPD_1^\star(\cdot\ ; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})$ is unimodal and the mode of the distribution is given by m where $h_1(m) = \eta_1$. If $\eta_1 \le 0$ then the probability density is strictly decreasing on $\mathbb{R}_+$.

(c) If $B > m$ is such that $\frac{g_1(B; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})}{g_1(m; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})} < \epsilon$ for some $\epsilon > 0$, then $P(d_1 > B \mid d_2, \ldots, d_p) < \epsilon$.

(d) Let M_{crit} be any positive number, then for all $d_1 > M_{crit}$,

$$g_1(d_1\,; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta}) \quad \le \quad K^\dagger_{n,p,M_{crit}}\, d_1^{\nu(n-1)/2}\, exp(\, -\nu(1-\eta_1)\, d_1), \tag{6.1}$$

where

$$K^\dagger_{n,p,M_{crit}} = \left[\frac{(p/4)^{\frac{n/2-1}{2}}}{\Gamma(n/2)\left\{\sqrt{M_{cric}}\ e^{-M_{crit}}\ I_{n/2-1}(M_{crit})\right\}}\right]^\nu.$$

Even though parts (a) and (b) of the above theorem follow immediately from Theorem 3, they are included here for completeness; all the properties play a crucial role in the construction of the sampling technique for $CCPD_j^\star$. The proof of part (c) is essentially an implication of the fact that the right tail of the distribution decays at an exponential rate. To show part (d) we have developed a nontrivial lower bound for ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$.

Remark 1. *The constant* $K^{\dagger}_{n,p,M_{crit}}$ *in part (d) of Theorem 9 converges to a finite constant as* M_{crit} *approaches infinity. It follows from the properties of the Bessel function that*

$$\lim_{M_{crit}\to\infty} \sqrt{M_{crit}} e^{-M_{crit}} I_{a-1}(M_{crit}) = \frac{1}{\sqrt{2\pi}}$$

for all $a \geq \frac{3}{2}$. *Hence for larger values of* M_{crit}, *the value of* $K^{\dagger}_{n,p,M_{crit}}$ *approaches* $\left[\frac{\sqrt{2\pi}(p/4)^{\frac{n/2-1}{2}}}{\Gamma(n/2)}\right]^{\nu}$, *a nonzero finite constant depending on* n, p, ν.

Note that the ratio $g_1(B; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})/g_1(m; \boldsymbol{d}^{(-1)}, \nu, \boldsymbol{\eta})$, mentioned in part (c), is free of the intractable normalizing constants of the distribution. Therefore, the numerical computation of the ratio is possible as long as we can compute the corresponding ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$. Using Theorem 9, we develop an accept-reject sampling algorithm that can generate samples from $CCPD^{\star}_j$ with high acceptance probability. The detailed construction of the sampler is provided next. We conclude this section with a description of an efficient procedure for computing the ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ constant.

6.1 A Rejection Sampler for the $CCPD^{\star}_j$ Distribution

We now describe a rejection sampling procedure from the conditional distribution of $(d_1 \mid (d_2, \cdots, d_p))$ when $\boldsymbol{d} \sim CCPC(\cdot; \nu, \boldsymbol{\eta})$ for some $\nu > 0$ and $\max_{1\leq j\leq p} \eta_j < 1$. Here $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$. Let m be the mode of the conditional distribution, $g_1(\cdot) := g(\cdot\,; \nu, \boldsymbol{\eta} \mid (d_2, \ldots, d_p))$, of the variable d_1 given $(d_2, \ldots, d_p)$ when $\eta_1 > 0$. In case $\eta_1 \leq 0$, we set m to be 0. Using the properties of the conditional distribution described in Theorem 9, we compute a critical point M_{crit} such that $P(d_1 > M_{crit} \mid (d_2, \cdots, d_p), \{X_j\}_{j=1}^{N}) < \epsilon$. Here we have chosen $\epsilon = 0.0001$.

To construct a proposal density $\overline{g}_1(x)$, we employ two different strategies, one for the bounded interval $(0, M_{crit}]$ and the other using Theorem 9 to tackle the tail, (M_{crit}, ∞), of the support of the conditional posterior distribution of d_1.

The procedure is as follows. Let $\delta = M_{crit}/N_{bin}$ where N_{bin} is the total number of partitions of the interval $(0, M_{crit}]$. Consider $k = ([m/\delta] + 1)$ where $[m/\delta]$ denotes the greatest integer less than or equal to m/δ. Now define the function

$$\begin{aligned}\overline{g}_1(x) \quad := \quad & \sum_{j=1}^{k-1} g_1(j\,\delta)\, \mathbb{I}_{((j-1)\delta, j\delta])}(x) + g_1(m)\mathbb{I}_{((k-1)\delta, k\delta])}(x) \\ & + \sum_{j=k+1}^{N_{bin}} g_1((j-1)\,\delta)\, \mathbb{I}_{(((j-1)\delta, j\delta])}(x) \\ & + K^{\dagger}_{n,p,M_{crit}}\, d_1^{\nu(n-1)/2} exp(\,-\nu(1-\eta_1)\, d_1) \mathbb{I}_{(M_{crit},\infty))}(x), \qquad (6.2)\end{aligned}$$

where $K^{\dagger}_{n,p,M_{crit}}$ is as defined in part (d) of Theorem 9.

From Theorem 9 it follows that $\overline{g}_1(x) \geq g_1(x)$ for all $x > 0$ as $g_1(\cdot)$ is a unimodal log-concave function with maxima at m. We consider,

$$q_j = \begin{cases} \delta\, g_1(j\delta) & \text{if } 1 \leq j < \left[\frac{m}{\delta}\right] + 1, \\ \delta\, g_1(m) & \text{if } j = \left[\frac{m}{\delta}\right] + 1, \\ \delta\, g_1((j-1)\delta) & \text{if } \left[\frac{m}{\delta}\right] + 1 < j \leq N_{bin}, \\ K^{\dagger}_{n,p,M_{crit}} \frac{\Gamma\left(\frac{(\nu(n-1)+2)}{2}, M\nu(1-\eta_1)\right)}{[\nu(1-\eta_1)]^{\nu(n-1)/2+1}} & \text{if } j = N_{bin}+1, \end{cases}$$

where $\Gamma\left(\frac{(\nu(n-1)+2)}{2}, M_{crit}\nu(1-\eta_1)\right)$ denotes the upper incomplete gamma function. For the case where M_{crit} tends to ∞ (see Remark 1) the constant $K^{\dagger}_{n,p,M_{crit}}$ approaches a finite constant, whereas $\Gamma\left(\frac{(\nu(n-1)+2)}{2}, M_{crit}\nu(1-\eta_1)\right)$ monotonically decreases to zero. Therefore, the positive constant $q_{N_{bin}+1}$ can be made arbitrary close to zero by choosing a suitably large value for M_{crit} when the value of n, p, ν, η_1 are fixed. Note that the quantities $\{q_j\}_{j=1}^{N_{bin}+1}$ may not add up to 1, therefore we construct the corresponding set of probabilities, $\{p_j\}_{j=1}^{N_{bin}+1}$ where $p_j = q_j / \sum_{j=1}^{N_{bin}+1} q_j$ for $j = 1, 2, \cdots, N_{bin}+1$. The following algorithm lists the steps involved in generating a sample from the distribution corresponding to the kernel $g_1(\cdot)$.

Algorithm 3 Steps for the rejection sampler for $CCPD^{\star}_{j}$.

1: Sample Z from the discrete distribution with the support $\{1, 2, \ldots, (N_{bin}+1)\}$ and corresponding probabilities $\{p_j\}_{j=1}^{N_{bin}+1}$,
2: **if** $Z \leq N_{bin}$ **then**
3: Sample $y \sim$ Uniform $((Z-1)\,\delta, Z\delta)$,
4: **else** Sample $y \sim$ TruncatedGamma$\left(\text{shape} = \frac{\nu(n-1)+2}{2},\ \text{rate} = \nu(1-\eta_1), \text{support} = (M_{crit}, \infty)\right)$
5: **end if**
6: Sample $U \sim$ Uniform $(0, 1)$,
7: **if** $U \leq \frac{g_1(y)}{\overline{g}_1(y)}$ **then**
8: Accept y as a legitimate sample from $g_1(\cdot)$
9: **else** Go to Step 1
10: **end if**

Figure 2 shows a typical example of the function $g_1(x)$ and the corresponding $\overline{g}_1(x)$. The blue curve represents the unnormalized density g_1. The black curve and the red curve after M_{crit} constitutes the function $\overline{g}_1$ (defined in (6.2)). Note that the red curve after the point M_{crit} represents the last term (involving $K^{\dagger}_{n,p,M_{crit}}$) in the summation formula in (6.2). In Figure 2(a), the values of δ and M_{crit} are set such that the key components of g_1 and $\overline{g}_1(x)$ are easy to discern. On the other hand, Figure 2(b) displays the plot of $\overline{g}_1(x)$ when recommended specification of M_{crit} and δ are used.

The choice of N_{bin} plays a crucial role in the algorithm and is required to be determined before constructing the proposal density for the accept-reject algorithm. Note

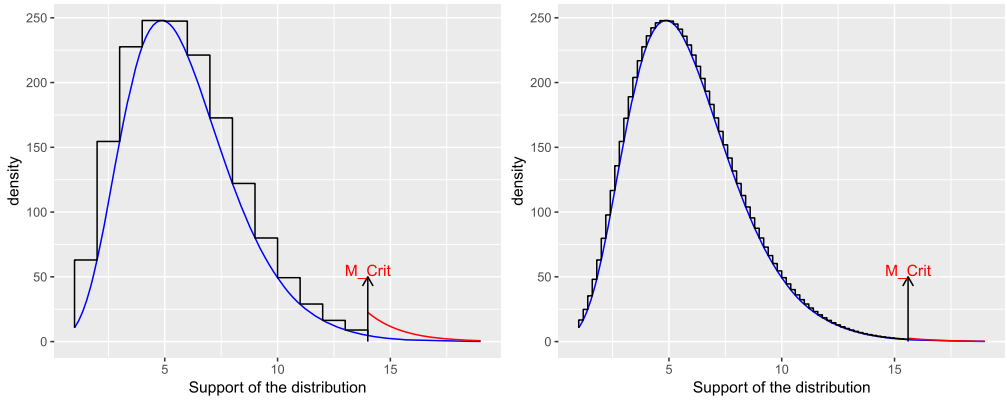


Figure 2: The blue curves represent g_1, the unnormalized density of $CCPD_1^\star$ distributions. The black curve and the red curve after M_{crit} constitutes the function $\overline{g}_1$, the proposal density for the accept reject algorithm. The panel(a) displays the key aspects of the densities while panel(b) shows the proposal density when recommended specifications of M_{crit} and δ are used.

that N_{bin} and δ are interconnected. If one is specified, the value of the other can be determined. We decide to choose the parameter δ and compute the corresponding N_{bin}. In the case where the concentration parameter is high, a finer partition of the proposal histogram (smaller value of δ) is required to keep the acceptance rate of the algorithm high. Based on our empirical results, we recommend selecting δ to be of the order of $\frac{1}{\sqrt{\nu}}$. The acceptance probability remains stable across different choices of ν when the value δ is set accordingly (see Figure 3). The estimated acceptance probabilities, used in Figure 3, were calculated based on 10000 Monte Carlo samples for each value of ν varied from 1 to 100. The relationship between N_{bin} and δ and ν is presented in Table 1.

Finally, successful implementation of the sampling algorithm developed in this subsection requires the computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$, a key step for the computation of $g_1(\cdot)$. In Section 6.2 we discuss the procedure that we have adopted to compute ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$.

6.2 Computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$

We first describe an efficient and reliable computational procedure to compute the function ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ when the argument matrix D is of dimension 2×2. The procedure is relevant to many applications considered in the field (Downs et al., 1971; Downs, 1972; Jupp and Mardia, 1979, 1980; Mardia and Khatri, 1977; Mardia et al., 2007; Mardia and Jupp, 2009; Chikuse, 1991a,b, 1998, 2003; Sei et al., 2013; Lin et al., 2017). We

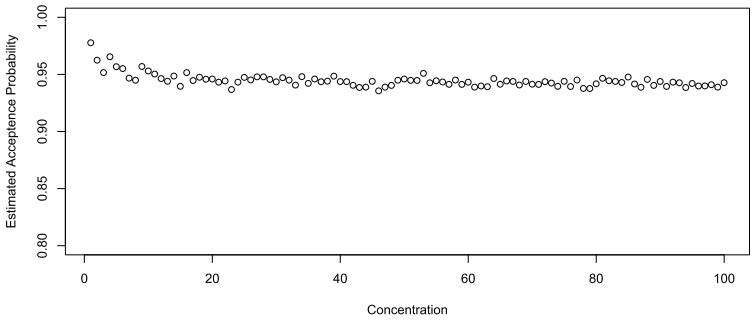


Figure 3: Estimated acceptance probability of the sampling algorithm when the value of the concentration parameter varies from 1 to 100. The parameter δ is chosen to be reciprocal of $\sqrt{\nu}$.

emphasize that the computational procedure described below is applicable for analyzing data on $\mathcal{V}_{n,2}$ for all $n \geq 2$.

Consider the representation developed in Muirhead (1975) for the Hypergeometric function of a matrix argument

$$ {}_0F_1\left(c, D\right) = \sum_{k=0}^{\infty} \frac{d_1^k d_2^k}{\left(c-\frac{1}{2}\right)_k \left(c\right)_{2k} k!} \, {}_0F_1\left(c+2k, d_1+d_2\right), \tag{6.3} $$

where D is a 2×2 diagonal matrix with diagonal elements $d_1 > 0, d_2 > 0$. From Butler and Wood (2003) (see page 361), it can be seen that,

$$ {}_0F_1\left(c+2k, d_1+d_2\right) = \frac{\Gamma\left(c+2k\right)}{\left(\sqrt{d_1+d_2}\right)^{(c+2k-1)}} I_{c+2k-1}\left(2\sqrt{d_1+d_2}\right), \tag{6.4} $$

where $I_{c+2k-1}(\cdot)$ is the modified Bessel function of the first kind with order $(c+2k-1)$. Hence from (6.3) and (6.4), we get that

$$ \begin{aligned} {}_0F_1\left(c, D\right) &= \sum_{k=0}^{\infty} \frac{d_1^k d_2^k}{\left(c-\frac{1}{2}\right)_k \left(c\right)_{2k} k!} \frac{\Gamma\left(c+2k\right) \;\; I_{c+2k-1}\left(2\sqrt{d_1+d_2}\right)}{\left(\sqrt{d_1+d_2}\right)^{(c+2k-1)}} \\ &= \sum_{k=0}^{\infty} A_k, \end{aligned} \tag{6.5} $$

where $A_k = \frac{\Gamma(c-.5)\Gamma(c)}{\Gamma(c+k-.5)k!} \frac{(d_1 d_2)^k}{\left(\sqrt{d_1+d_2}\right)^{(c+2k-1)}} \; I_{c+2k-1}\left(2\sqrt{d_1+d_2}\right)$. Note that

$$ \begin{aligned} \frac{A_{k+1}}{A_k} &= \frac{\Gamma(c+k-.5)k!}{\Gamma(c+k+.5)(k+1)!} \frac{I_{c+2k+1}\left(2\sqrt{d_1+d_2}\right)}{I_{c+2k-1}\left(2\sqrt{d_1+d_2}\right)} \frac{d_1 d_2}{(d_1+d_2)} \\ &\leq \frac{4 d_1 d_2}{(2c+2k-1)(2k+2)(2k+c)(2k+2c+1)}, \end{aligned} \tag{6.6} $$

ν	δ	Estimated Acceptance probability	N_{bin}
1	1	0.95813	42
1	0.5	0.977517	85
1	0.333333	0.984155	127
1	0.2	0.988924	212
1	0.1	0.996314	425
1	0.05	0.998104	851
3	0.5	0.952835	27
3	0.333333	0.963206	40
3	0.2	0.977326	67
3	0.1	0.988924	135
3	0.05	0.995124	271
5	1	0.885818	3
5	0.5	0.941886	7
5	0.333333	0.960246	10
5	0.2	0.973994	17
5	0.1	0.989218	35
5	0.05	0.993246	71

Table 1: Values of the N_{bin}, δ and acceptance probability for algorithm to generate values from $CCPD_j(\eta, \nu)$ for $\nu = 1, 3, 5$.

where the last inequality follows from $I_{\nu+1}(x)/I_\nu(x) < \frac{x}{2(\nu+1)}$ for $x > 0, \nu > -1$ (see page 221 in Ifantis and Siafarikas (1990)). For fixed values of d_1, d_2 we can find M such that $A_M \leq \epsilon$ and $M^4 \geq (d_1\ d_2)/(4\epsilon_1)$ for some $\epsilon_1 < \frac{1}{2}$ and a predetermined error bound ϵ. For such a choice of M, if k is any integer such that $k \geq M$, then

$$\begin{aligned}
\frac{A_{k+1}}{A_k} &\leq \frac{4d_1d_2}{(2c+2k-1)(2k+2)(2k+c)(2k+2c+1)} \\
&\leq \frac{4d_1d_2}{(2c+2M-1)(2M+2)(2M+c)(2M+2c+1)} \\
&\leq \left(\frac{d_1d_2}{4M^4}\right)\left\{\frac{16M^4}{(2c+2M-1)(2M+2)(2M+c)(2M+2c+1)}\right\} \\
&\leq \left(\frac{d_1d_2}{4M^4}\right)\left\{\frac{M^4}{(M+\frac{2c-1}{2})(M+1)(M+\frac{c}{2})(M+\frac{2c+1}{2})}\right\} \\
&\leq \epsilon_1, \qquad (6.7)
\end{aligned}$$

where the last inequality follows due to the fact that $M^4 \leq (M+\frac{2c-1}{2})(M+1)(M+\frac{c}{2})(M+\frac{2c+1}{2})$ as $c > \frac{1}{2}$. Hence from (6.5) we get that

$$|{}_0F_1\,(c, D) - \sum_{k=0}^{M} A_k| = \sum_{k=M+1}^{\infty} A_k \leq A_M \sum_{k=M+1}^{\infty} \epsilon_1^{k-M} \leq \frac{\epsilon\ \epsilon_1}{1-\epsilon_1} < \epsilon. \qquad (6.8)$$

Consequently, for a given value of the matrix D and an error level ϵ, we can select M accordingly, so that ${}_0F_1(c, D)$ is approximated as

$$ {}_0F_1(c, D) \approx \sum_{k=0}^{M} \frac{d_1^k d_2^k}{\left(c - \frac{1}{2}\right)_k (c)_{2k}\, k!} \frac{\Gamma(c+2k)\ \ I_{c+2k-1}\left(2\sqrt{d_1 + d_2}\right)}{\left(\sqrt{d_1 + d_2}\right)^{(c+2k-1)}}, \tag{6.9} $$

where the error in the approximation is at most ϵ.

In the case when the matrix D is of dimension $p \times p$ with $p > 2$, we rely on the computational technique developed in Koev and Edelman (2006). Development of efficient computational schemes for the hyper geometric function of a matrix argument in general dimension is an active area of research (Gutiérrez et al., 2000; Koev and Edelman, 2006; Nagar et al., 2015; Pearson et al., 2017). In principle, the theoretical framework developed in this article integrated with the general computation scheme specified in Koev and Edelman (2006) can handle data on $\mathcal{V}_{n,p}$ for arbitrary integers $n \geq p \geq 2$, but the results from the combined procedure may lack precision as it inherits the limitations of the algorithm in Koev and Edelman (2006) (see page 835 in Koev and Edelman (2006)). In the following remark we specify the assumptions under which the combined procedure can be applied effectively.

Remark 2. *The algorithm developed in Koev and Edelman (2006) is a general procedure for computing ${}_pF_q(\cdot)$ for arbitrary integers $p, q \geq 0$. Naturally, the algorithm applies to ${}_0F_1$ which is the object of focus in the current context. Due to its generality, the computational scheme has certain limitations. In particular, it requires appropriate specification of a "tuning parameter" that can not be determined in an automated manner. However, from an empirical exploration of the procedure, we observed that the corresponding outputs can be quite robust. Particularly, the output was found to stabilize after a certain point (we will call this the "stabilization point") when the value of the tuning parameter was gradually increased. For the case of $p = 2$, if the tuning parameter is specified to be larger than the stabilization point, the output from Koev and Edelman (2006) is very close to the true value, as determined by our arbitrary precision algorithm. Extrapolating to $p \geq 3$, we presume that the true value of the corresponding hyper geometric function will be close to the output of Koev and Edelman (2006) if the tuning parameter is set larger than the "stabilization point". As the "stabilization point" is observed to be larger for larger values of D, we can set the value of the tuning parameter to a single prespecified number for an entire analysis only if we assume that the diagonal elements of the matrix D are bounded above by a prespecified finite number. Under this assumption, we can rely on Koev and Edelman (2006) for the analysis of data on $\mathcal{V}_{n,p}$, $n \geq p \geq 3$. In that case, the combination of our theoretical framework and the algorithm for the computation of the hypergeometric function from Koev and Edelman (2006) would work effectively for practical applications (see Simulation Section 7.2).*

In contrast, the procedure to compute ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ that we have developed, though targeted towards a specific case, has a theoretical guarantee for a desired level of precision of its output. Since many statistical applications, as mentioned earlier, are about analyzing data on $\mathcal{V}_{n,2}$, the computation procedure we have designed specifically for $\mathcal{V}_{n,2}$ has its own merit.

7 Simulation

To evaluate the performance of the procedure presented in the previous sections, we performed simulation experiments. We considered two different setups. In the first, we analyzed simulated datasets in $\mathcal{V}_{n,p}$ where we varied n to assess its effect on the posterior estimation efficiency. Here, the value of p was fixed at 2 and the computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ developed in Section 6.2 was utilized. In the second setup, we analyzed data on $\mathcal{V}_{n,p}$ to demonstrate the generic applicability of our framework by setting $p = 3$, $n = 5$. Here, we used the procedure in Koev and Edelman (2006) to calculate the value ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$.

7.1 Simulation Setup ($p = 2$)

We present results from experiments with simulated data where we varied the dimension of the Stiefel manifold, n, across a range of values. The objective of this simulation study was to see how the error rates varied with the dimension n. Specifically, we generated 3000 observations using $\mathcal{ML}$ distribution on $\mathcal{V}_{3,2}$, $\mathcal{V}_{5,2}$, $\mathcal{V}_{10,2}$, and $\mathcal{V}_{15,2}$. These correspond to the Stiefel Manifolds with dimension $[n = 3, p = 2]$, $[n = 5, p = 2]$, $[n = 10, p = 2]$, and $[n = 15, p = 2]$, respectively. We generated 50 datasets for each simulation setting using the algorithm mentioned in Hoff (2009). In order to generate data for each dataset we fixed the parameters M and V to the canonical orthogonal vectors of appropriate dimension and generated two entries of the parameter D from two independent gamma distributions.

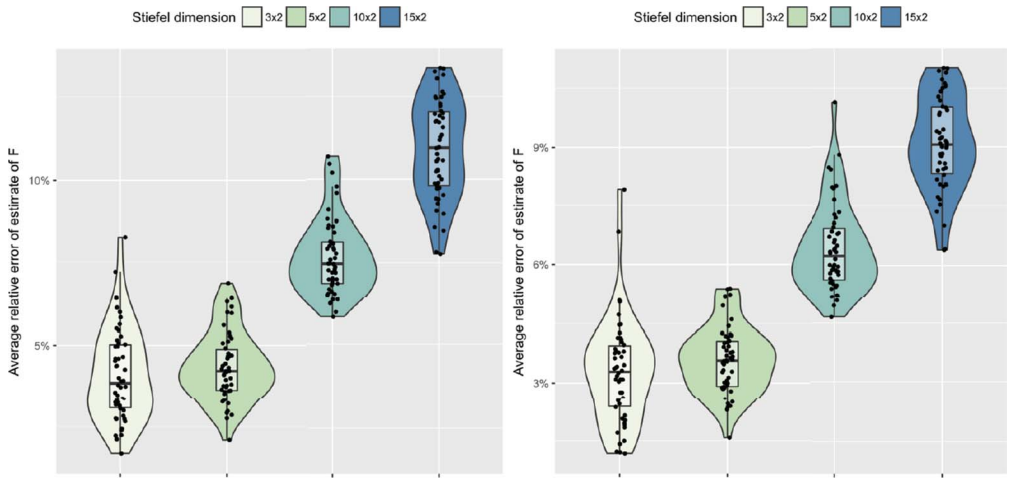
We ran posterior inference for each of these datasets using 3000 MCMC samples with an initial 1000 samples as burn-in. We used the posterior mean of the parameter F as the point estimate $\widehat{F}$. Finally we assessed our performance by computing the relative error for the estimate of $F_{true} = M_{true} D_{true} V_{true}^T$. We define the relative error as:

$$\frac{\|\widehat{F} - F_{true}\|}{\|F_{true}\|},$$

where $\|\cdot\|$ denotes the matrix Frobenious norm. Figure 4 shows the average relative error with the corresponding standard deviation of estimation for $\mathcal{V}_{3,2}$, $\mathcal{V}_{5,2}$, $\mathcal{V}_{10,2}$, and $\mathcal{V}_{15,2}$ for $N = 2000$ (panel (a)) and for $N = 3000$ (panel (b)). The average relative errors do not seem to exceed 11% and 9% for $N = 2000$ and 3000, respectively even with the dimension as high as 15. The error rate tends to increase with higher dimension, i.e., value of n. Also, we investigated the relationship with the total sample size and found these error rates to decrease with larger sample sizes. For example, the reduction in average relative error rate for $n = 5$ and $N = 2000$ is around 2%. Overall, these results demonstrate the robustness of our inference procedure.

7.2 Simulation Setup ($p > 2$)

Having demonstrated the efficiency of our method for a range of values of n with $p = 2$, we now present an example of a generalized simulation scenario for $p > 2$. Here we



(a) with 2000 data points (b) with 3000 data points

Figure 4: Relative error of $\widehat{F}$ for matrices with different dimensions.

use the procedure in Koev and Edelman (2006) to numerically approximate the value of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ where D is a $p \times p$ dimensional matrix with $p > 2$ (see Remark 2). Through the entire simulation we fixed the tuning parameter required in the computation of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ to a large prespecified value. Here we give a specific example with $n = 5$ and $p = 3$. We generated 50 datasets of 500 observations each using the $\mathcal{ML}$ distribution with different parameters, on $\mathcal{V}_{5,3}$. We then ran posterior inference for each of these datasets using 1100 MCMC samples with an initial 100 sample burn-in. We used the posterior mean of the parameter F as before as the estimate of the true parameter F. Using the same metric we computed the average relative error of the estimation (Figure 5). We observed that our sampling algorithm for d_i $(i = 1, 2, 3)$ runs with a very low rejection rate. As can be seen in Figure 5, the average relative errors do not exceed 3%, demonstrating the general applicability of our framework beyond $p = 2$.

Codes for the algorithms are available at https://github.com/ssra19/Stiefel_Bayes.git.

8 Application

Finally, to showcase the methodology developed in this paper, we analyzed the vectorcardiogram (VCG) dataset discussed in Downs et al. (1971). The dataset contains vectorcardiograms of 56 boys and 42 girls aged between 2 and 19 years. Individuals in the dataset are partitioned into four groups: groups 1 and 2 consist of boys aged between 2–10 and 11–19 years, while groups 3 and 4 consist of girls aged between 2–10 and 11–19 years. Each sample contains vectorcardiograms acquired using two different

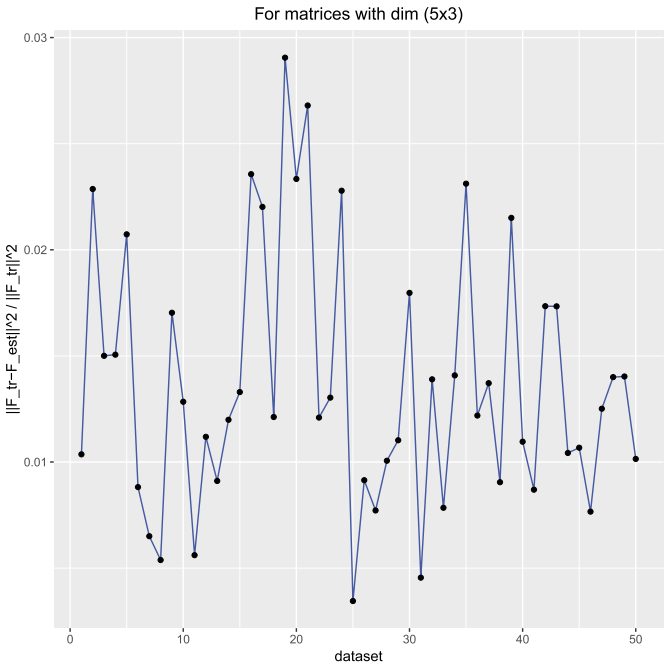


Figure 5: Average relative error for datasets on $\mathcal{V}_{5,3}$.

measurement systems, the Frank lead system (Frank, 1956; Downs et al., 1971) and the McFee lead system (Downs et al., 1971). Here, we restrict ourselves to groups 1 and 3 and measurements acquired using the McFee lead system. For each individual sample, we considered the pair of orthogonal vectors that provides the orientation of the "QRS loop" (Downs et al., 1971) in $\mathbb{R}^3$. Each orientation in the sample is defined by a 3×2 matrix with orthonormal columns, i.e., an element in $\mathcal{V}_{3,2}$. Additional details regarding the measurements, data structures, and data processing can be found in Downs et al. (1971).

8.1 MCMC Convergence Diagnostics

We ran several MCMC convergence diagnostic tests for the MCMC samples from the posterior of $F = MDV^T$, which is the natural parameter of the Matrix Langevin distribution. The parameter F uniquely identifies and is uniquely identified by the parameters M, D, V. Moreover the elements of the matrix M and V are interrelated whereas the components of F are not thus constrained. We therefore focused the diagnostics on F and studied its estimation accuracy. As notation, $F_{i,j}$ denotes the $[i, j]$-th element of F. We first ran convergence diagnostics based on potential scale reduction factor (PSRF) (Gelman et al., 1992). We ran the MCMC procedure three times with different random seeds for 10,000 MCMC iterations with a 1000 sample burn-in. The PSRF is a weighted sum of within-chain and between-chain variances. The calculated PSRF was

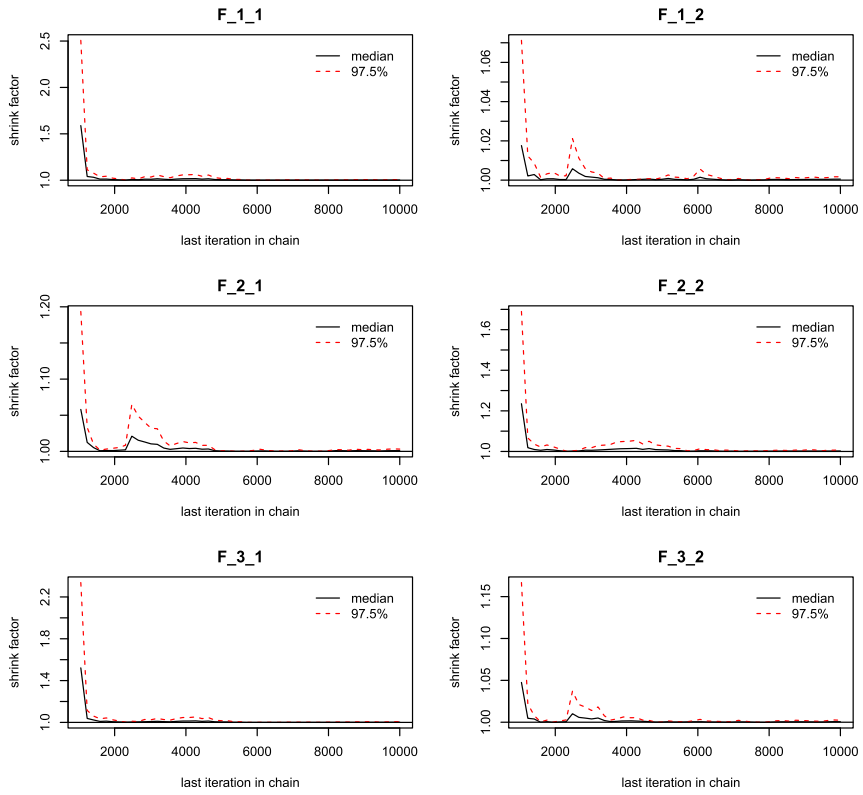
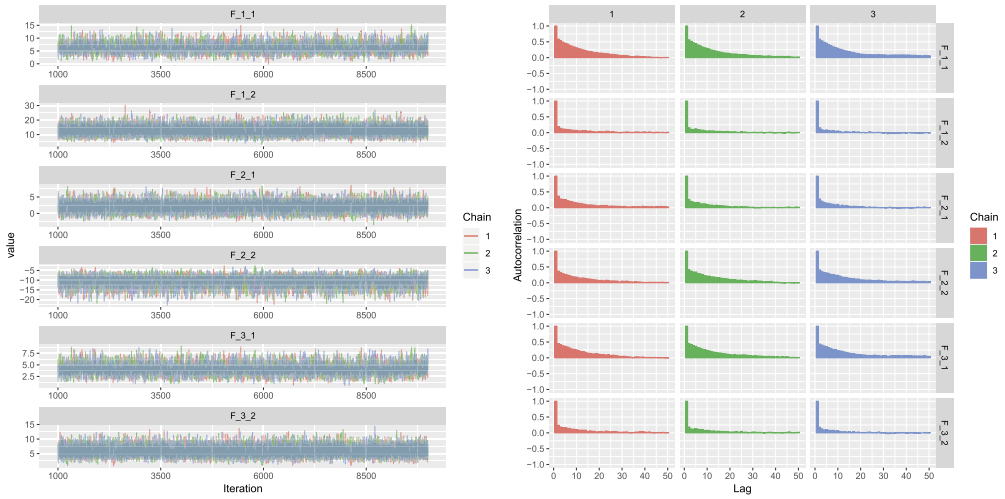


Figure 6: PSRF for all six components of posterior samples of F.

1.00 with an upper confidence bound 1.01, indicating no evidence of lack of convergence. We show how the PSRF changed with the iterations in Figure 6 for all components of F. We also calculated a multivariate potential scale reduction factor (MPSRF) that was proposed by Gelman and Brooks (Brooks and Gelman, 1998). The calculated MPSRF was 1.01, also confirming that there was no lack of convergence. The log-likelihood is yet another measure representative of the multi-dimensional parameters. In this case too, the calculated PSRF for log-likelihood was 1.0 with an upper confidence bound 1.0, indicating no evidence of lack of convergence. Finally, we calculated the Heidelberg and Welch (HW) diagnostic (Heidelberger and Welch, 1981, 1983) which is a test statistic based on the Cramer-von Mises test statistic to accept or reject the null hypothesis that the MC is from a stationary distribution. This diagnostic has two parts and the MC chain for F passed both the Stationarity and Halfwidth Mean tests. This test too, then, showed no evidence for lack of convergence.

Figures 7(a), 7(b) and 8 show the traceplots, autocorrelations and densities of different components of the posterior samples of F from the three runs, respectively. Notably, the densities of all the components of F are unimodal, confirming convergence.



(a) traceplots (b) autocorrelations

Figure 7: Traceplots and autocorrelations of all six components of posterior samples of F from three runs.

8.2 Parameter Estimation

We modeled the VCG dataset using $\mathcal{ML}$ distributions on $\mathcal{V}_{3,2}$. There were 28 and 17 observations in groups 1 and 3, respectively. We assumed that each i.i.d. observation in group 1 follows a $\mathcal{ML}$ distribution with parameters $M_{group1}, \boldsymbol{d}_{group1}$ and V_{group1}, and likewise, i.i.d. observations in group 3 follow a $\mathcal{ML}$ distribution with parameters $M_{group3}, \boldsymbol{d}_{group3}$ and V_{group3}. We used the uniform improper prior for estimation of the parameters related to both groups (see Section 4). From (5.4), we note that the posterior distributions of $(M_{group1}, \boldsymbol{d}_{group1}, V_{group1})$ and $(M_{group3}, \boldsymbol{d}_{group3}, V_{group3})$ given the data are

$$JCPD\left(\cdot\,;\,28,\,\overline{W}_{group1}\right) \text{ and } JCPD\left(\cdot\,;\,17,\,\overline{W}_{group3}\right) \text{ where}$$

$$\overline{W}_{group1} = \begin{bmatrix} 0.687 & 0.576 \\ 0.551 & -0.737 \\ 0.122 & 0.142 \end{bmatrix} \text{ and } \overline{W}_{group3} = \begin{bmatrix} 0.682 & 0.585 \\ 0.557 & -0.735 \\ 0.125 & 0.055 \end{bmatrix}$$

are the sample means of the observations in groups 1 and 3, respectively. We verified the spectral norm condition in Theorem 1 for the posterior distributions to be well defined; we found $\left\|\overline{W}_{group1}\right\|_2 = 0.946$ and $\left\|\overline{W}_{group3}\right\|_2 = 0.941$.

Using Theorem 3, we can infer that the above-mentioned posterior distributions have unique modes. Also from Theorem 3 we can compute the posterior mode and they were

$$\widehat{M}_{group1} = \begin{bmatrix} -0.650 & 0.733 \\ 0.743 & 0.668 \\ -0.157 & 0.127 \end{bmatrix}, \widehat{\boldsymbol{d}}_{group1} = \begin{bmatrix} 16.329 \\ 5.953 \end{bmatrix}, \widehat{V}_{group1} = \begin{bmatrix} -0.059 & 0.998 \\ -0.998 & -0.059 \end{bmatrix}.$$

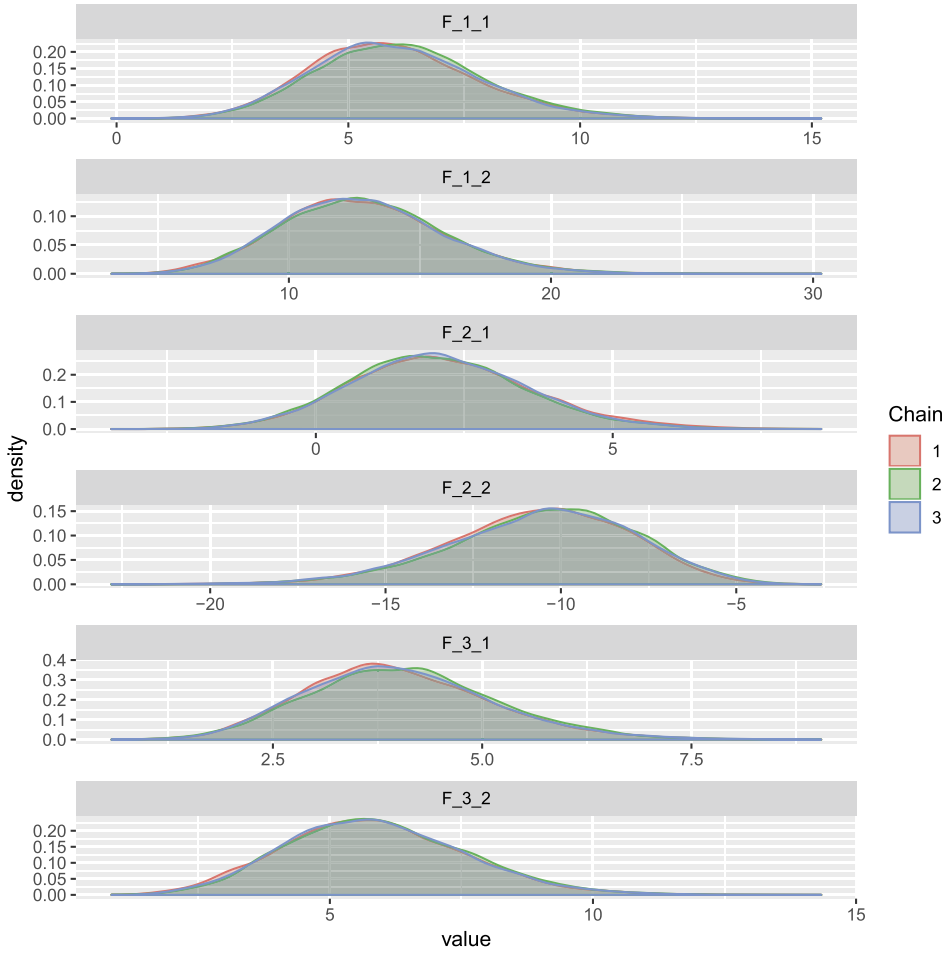


Figure 8: Densities of all six components of posterior samples of F from three runs.

Similarly, we can compute the posterior mode for the parameters of group 3 (not reported here). To estimate the posterior mean for the parametric functions

$$F_{group1} = M_{group1} D_{group1} V_{group1}^T \text{ and } F_{group3} = M_{group3} D_{group3} V_{group3}^T,$$

we ran the MCMC based posterior inference procedure described in Section 6 to generate MCMC samples from each of the posterior distribution.

For group 1, the posterior mean for the parametric function $F_{group1} = M_{group1} D_{group1} V_{group1}^T$ was

$$\hat{\hat{F}}_{group1} = \begin{bmatrix} 5.183 & 9.086 \\ 3.583 & -10.996 \\ 0.919 & 2.221 \end{bmatrix}, \; SD(\hat{\hat{F}}_{group1}) = \begin{bmatrix} 1.527 & 2.354 \\ 1.475 & 2.665 \\ 0.596 & 0.898 \end{bmatrix},$$

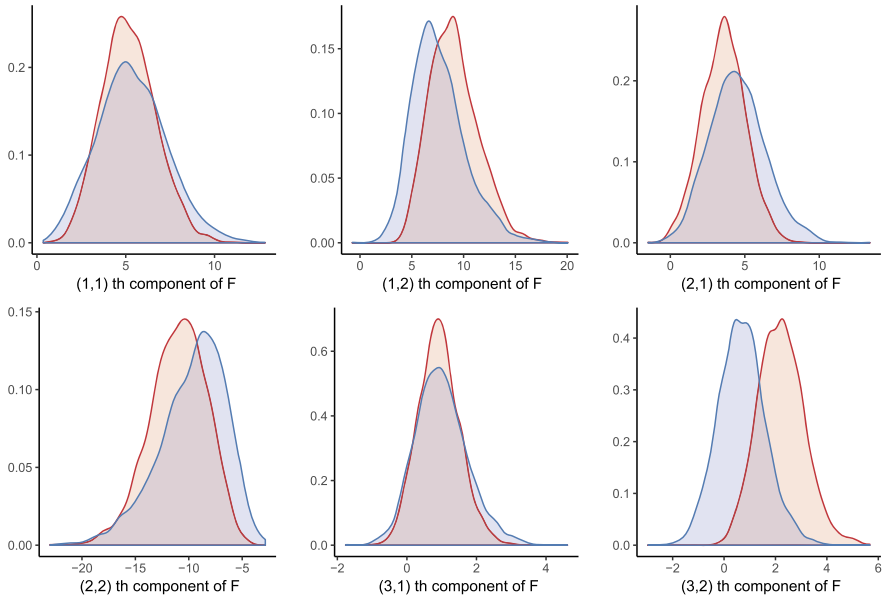


Figure 9: Estimated posterior density for the parameter F. The estimated density for Group 1 and Group 3 are marked with Red and Blue lines respectively.

where the entries of the matrix $SD(\hat{\hat{F}}_{group1})$ provides the standard deviation for the corresponding entries of $\hat{\hat{F}}_{group1}$. From the MCMC samples, we also estimated the posterior density of each entry of F_{group1} and F_{group3}. Figure 9 shows the corresponding density plots. The estimates related to group 3 were

$$\hat{\hat{F}}_{group3} = \begin{bmatrix} 3.249 & 8.547 \\ 3.798 & -10.658 \\ 1.605 & 0.796 \end{bmatrix} \text{ and } SD(\hat{\hat{F}}_{group3}) = \begin{bmatrix} 1.263 & 2.123 \\ 1.359 & 2.624 \\ 0.603 & 0.83 \end{bmatrix}.$$

8.3 Hypothesis Testing

Finally, we conducted a two sample hypothesis test for comparing different data groups on the Stiefel manifold. We have chosen hypothesis testing as one of our demonstrations because a general two sample test that does not rely on asymptotics or on the concentration being very large or very small, has not been reported in the literature for data lying on the Stiefel manifold (Khatri and Mardia, 1977; Chikuse, 2012). The procedure described here is valid for finite sample sizes and does not require any additional assumptions on the magnitude of the parameters.

We considered the VCG dataset and carried out a test to compare the data group 1 against the data group 3, i.e.

$$H_0 : F_{group1} = F_{group3} \text{ vs } H_A : F_{group1} \neq F_{group3}.$$

To test the hypotheses in a Bayesian model selection framework, we considered two models $Model_0$ and $Model_1$. In $Model_0$, we assumed $M_{group1} = M_{group3}$, $\boldsymbol{d}_{group1} = \boldsymbol{d}_{group3}$, $V_{group1} = V_{group3}$ while in $Model_1$, we did not impose any structural dependencies between the parameters. We assumed the prior odds between the models to be 1 and computed the Bayes factor

$$B_{0,1} = \frac{P(Data \mid Model_0)}{P(Data \mid Model_1)},$$

where $Data$ denotes the combined data from both groups. Since an analytic form for the Bayes factor is not available in this case, we used an MCMC based sampling technique to estimate the Bayes factor. We used the empirical prior (see Section 4) with the choice of prior concentration set at 1 percentage of the corresponding sample size. We followed the procedure described in Section 6 to generate MCMC samples from each of the required posterior distribution. We used the harmonic mean estimator (HME) (Newton and Raftery, 1994) to estimate the marginal likelihoods required for computing the Bayes factor. It is well known that the HME may not perform well when using improper priors. Consequently, unlike in Section 8.2 where we focus on the parameter estimation, we use an informative prior for this part of the analysis. We observed that the HME estimator is stable for the current context. The estimate of $log(B_{01})$ was 51.994. Hence, we conclude that there is not enough evidence to favor $Model_1$ over $Model_0$.

9 Discussion and Future Directions

In this article, we have formulated a comprehensive Bayesian framework for analyzing data drawn from a $\mathcal{ML}$ distribution. We constructed two flexible classes of distributions, *CCPD* and *JCPD*, which can be used for constructing conjugate priors for the $\mathcal{ML}$ distribution. We investigated the priors in considerable detail to build insights into their nature, and to identify interpretations for their hyper-parameter settings. Finally, we explored the features of the resulting posterior distributions and developed efficient computational procedures for posterior inference. An immediate extension would be to expand the framework to mixtures of $\mathcal{ML}$ distributions, with applications to clustering of data on the Stiefel manifold.

On a related note, we observed that the tractability of the set of procedures proposed in this article depends crucially on one's capacity to compute the hypergeometric function ${}_0F_1\left(n/2, F^TF/4\right)$ as a function the matrix F. We were naturally led to a modified representation of ${}_0F_1\left(n/2, D^2/4\right)$ (see Section 2) as a function of a vector argument $\boldsymbol{d}$. We explored several properties of the function ${}_0F_1\left(n/2, D^2/4\right)$, that are applicable to research areas far beyond the particular problem of interest in this article. As a special note, we should highlight that we designed a tractable procedure to compute the hypergeometric function of a $n \times 2$ dimensional matrix argument. There are many applications in the literature (Mardia and Khatri, 1977; Jupp and Mardia, 1979; Chikuse, 1998, 2003; Lin et al., 2017) where the mentioned computational procedure of ${}_0F_1\left(\frac{n}{2}, \frac{D^2}{4}\right)$ can make a significant impact. As such, the manner in which we have approached this computation is entirely novel in this area of research and the procedure

is scalable to "high-dimensional" data, such as in diffusion tensor imaging. In the near future, we plan to further explore useful analytical properties of the hypergeometric function, and extend our procedure to build reliable computational techniques for the hyper-geometric function where the dimension of the matrix argument is $n \times p$ with $p \geq 3$.

Finally, there is scope for extending the newly proposed family of prior distributions to a larger class of Bayesian models involving more general densities on manifolds. The properties of the prior and posterior discovered can also be seamlessly generalized. The coming together of state-of-the-art Bayesian methods incorporating topological properties of the underlying space promises to be a rich area of research interest.

Supplementary Material

Supplementary material: Conjugate Priors and Posterior Inference for the Matrix Langevin Distribution on the Stiefel Manifold (DOI: 10.1214/19-BA1176SUPP; .pdf).

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press. MR2364186. doi: https://doi.org/10.1515/9781400830244. 875

Bhatia, R. (2009). *Positive definite matrices*, volume 24. Princeton university press. MR2284176. 876

Brooks, S. P. and Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7(4): 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 899

Butler, R. W. and Wood, A. T. (2003). "Laplace approximation for Bessel functions of matrix argument." *Journal of Computational and Applied Mathematics*, 155(2): 359–382. MR1984295. 875, 893

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA. MR1051420. 873, 876, 878

Chikuse, Y. (1991a). "Asymptotic expansions for distributions of the large sample matrix resultant and related statistics on the Stiefel manifold." *Journal of Multivariate Analysis*, 39(2): 270–283. MR1147122. doi: https://doi.org/10.1016/0047-259X(91)90101-7. 872, 873, 892

Chikuse, Y. (1991b). "High dimensional limit theorems and matrix decompositions on the Stiefel manifold." *Journal of Multivariate Analysis*, 36(2): 145–162. 872, 873, 892

Chikuse, Y. (1998). "Density estimation on the Stiefel manifold." *Journal of Multivariate Analysis*, 66(2): 188–206. MR1642465. doi: https://doi.org/10.1006/jmva.1998.1747. 872, 873, 892, 903

Chikuse, Y. (2003). "Concentrated matrix Langevin distributions." *Journal of Multivariate Analysis*, 85(2): 375–394. MR1983803. doi: https://doi.org/10.1016/S0047-259X(02)00065-9. 873, 876, 892, 903

Chikuse, Y. (2012). *Statistics on Special Manifolds*, volume 174. Springer Science & Business Media. 871, 872, 875, 876, 902

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity, and applications*. Elsevier. MR0954608. 880

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of Statistics*, 7(2): 269–281. MR0520238. 873, 876, 877, 887

Doss, C. R. and Wellner, J. A. (2016). "Mode-constrained estimation of a log-concave density." *arXiv preprint* arXiv:1611.10335. MR3290441. doi: https://doi.org/10.1214/14-SS107. 880

Downs, T., Liebman, J., and Mackay, W. (1971). "Statistical methods for vectorcardiogram orientations." *Vectorcardiography*, 2: 216–222. 873, 892, 897, 898

Downs, T. D. (1972). "Orientation statistics." *Biometrika*, 665–676. MR0345334. doi: https://doi.org/10.1093/biomet/59.3.665. 871, 872, 873, 875, 892

Edelman, A., Arias, T. A., and Smith, S. T. (1998). "The geometry of algorithms with orthogonality constraints." *SIAM Journal on Matrix Analysis and Applications*, 20(2): 303–353. MR1646856. doi: https://doi.org/10.1137/S0895479895290954. 875

Frank, E. (1956). "An accurate, clinically practical system for spatial vectorcardiography." *Circulation*, 13(5): 737–749. 898

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL. MR3235677. 878

Gelman, A., Rubin, D. B., et al. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4): 457–472. 898

Gross, K. I. and Richards, D. S. P. (1987). "Special functions of matrix argument. I. Algebraic induction, zonal polynomials, and hypergeometric functions." *Transactions of the American Mathematical Society*, 301(2): 781–811. MR0882715. doi: https://doi.org/10.2307/2000670. 875

Gross, K. I. and Richards, D. S. P. (1989). "Total positivity, spherical series, and hypergeometric functions of matrix argument." *Journal of Approximation Theory*, 59(2): 224–246. MR1022118. doi: https://doi.org/10.1016/0021-9045(89)90153-6. 872, 875

Gupta, R. D. and Richards, D. S. P. (1985). "Hypergeometric functions of scalar matrix argument are expressible in terms of classical hypergeometric functions." *SIAM Journal on Mathematical Analysis*, 16(4): 852–858. MR0793927. doi: https://doi.org/10.1137/0516064. 875

Gutiérrez, R., Rodriguez, J., and Sáez, A. (2000). "Approximation of hypergeometric functions with matricial argument through their development in series of zonal polynomials." *Electronic Transactions on Numerical Analysis*, 11: 121–130. MR1799027. 895

Heidelberger, P. and Welch, P. D. (1981). "A spectral method for confidence interval generation and run length control in simulations." *Communications of the ACM*, 24(4): 233–245. MR0611745. doi: https://doi.org/10.1145/358598.358630. 899

Heidelberger, P. and Welch, P. D. (1983). "Simulation run length control in the presence of an initial transient." *Operations Research*, 31(6): 1109–1144. 899

Herz, C. S. (1955). "Bessel functions of matrix argument." *The Annals of Mathematics*, 474–523. MR0069960. doi: https://doi.org/10.2307/1969810. 875

Hill, R. D. and Waters, S. R. (1987). "On the cone of positive semidefinite matrices." *Linear Algebra and its Applications*, 90: 81–88. MR0884112. doi: https://doi.org/10.1016/0024-3795(87)90307-7. 876

Hobert, J. P., Roy, V., and Robert, C. P. (2011). "Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modeling." *Statistical Science*, 26(3): 332–351. MR2918006. doi: https://doi.org/10.1214/11-STS365. 872

Hoff, P. D. (2009). "Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data." *Journal of Computational and Graphical Statistics*, 18(2): 438–456. MR2749840. doi: https://doi.org/10.1198/jcgs.2009.07177. 889, 896

Hornik, K. and Grün, B. (2013). "On conjugate families and Jeffreys priors for von Mises-Fisher distributions." *Journal of Statistical Planning and Inference*, 143(5): 992–999. MR3011309. doi: https://doi.org/10.1016/j.jspi.2012.11.003. 877

Hornik, K. and Grün, B. (2014). "movMF: An R package for fitting mixtures of von Mises-Fisher distributions." *Journal of Statistical Software*, 58(10): 1–31. 872

Ibragimov, I. A. (1956). "On the composition of unimodal distributions." *Theory of Probability & Its Applications*, 1(2): 255–260. MR0087249. 880

Ifantis, E. and Siafarikas, P. (1990). "Inequalities involving Bessel and modified Bessel functions." *Journal of Mathematical Analysis and Applications*, 147(1): 214–227. MR1044696. doi: https://doi.org/10.1016/0022-247X(90)90394-U. 894

James, A. T. (1964). "Distributions of matrix variates and latent roots derived from normal samples." *The Annals of Mathematical Statistics*, 475–501. MR0181057. doi: https://doi.org/10.1214/aoms/1177703550. 875

James, I. M. (1976). *The Topology of Stiefel Manifolds*, volume 24. Cambridge University Press. MR0431239. 871

Jupp, P. and Mardia, K. (1980). "A general correlation coefficient for directional data and related regression problems." *Biometrika*, 163–173. MR0570518. doi: https://doi.org/10.1093/biomet/67.1.163. 872, 873, 892

Jupp, P. E. and Mardia, K. V. (1979). "Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions." *The Annals of Statistics*, 599–606. MR0527495. 872, 873, 887, 892, 903

Khare, K., Pal, S., Su, Z., et al. (2017). "A Bayesian approach for envelope models." *The Annals of Statistics*, 45(1): 196–222. MR3611490. doi: https://doi.org/10.1214/16-AOS1449. 878

Khatri, C. and Mardia, K. (1977). "The von Mises-Fisher matrix distribution in orientation statistics." *Journal of the Royal Statistical Society. Series B (Methodological)*, 95–106. MR0494687. 872, 875, 876, 902

Koev, P. and Edelman, A. (2006). "The efficient evaluation of the hypergeometric function of a matrix argument." *Mathematics of Computation*, 75(254): 833–846. MR2196994. doi: https://doi.org/10.1090/S0025-5718-06-01824-2. 873, 875, 895, 896, 897

Kristof, W. (1969). "A theorem on the trace of certain matrix products and some applications." *ETS Research Report Series*, 1969(1). MR0271125. doi: https://doi.org/10.1016/0022-2496(70)90037-4. 881, 882

Lin, L., Rao, V., and Dunson, D. (2017). "Bayesian nonparametric inference on the Stiefel manifold." *Statistica Sinica*, 27: 535–553. MR3674685. 871, 872, 873, 875, 892, 903

Lui, Y. and Beveridge, J. (2008). "Grassmann registration manifolds for face recognition." *Computer Vision–ECCV 2008*, 44–57. 871

Mardia, K. and Khatri, C. (1977). "Uniform distribution on a Stiefel manifold." *Journal of Multivariate Analysis*, 7(3): 468–473. MR0445720. doi: https://doi.org/10.1016/0047-259X(77)90087-2. 871, 872, 873, 892, 903

Mardia, K. V. and Jupp, P. E. (2009). *Directional Statistics*, volume 494. John Wiley & Sons. MR1828667. 871, 872, 873, 875, 876, 892

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). "Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data." *Biometrics*, 63(2): 505–512. MR2370809. doi: https://doi.org/10.1111/j.1541-0420.2006.00682.x. 873, 892

Muirhead, R. J. (1975). "Expressions for some hypergeometric functions of matrix argument with applications." *Journal of Multivariate Analysis*, 5(3): 283–293. MR0381137. doi: https://doi.org/10.1016/0047-259X(75)90046-9. 875, 893

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons. 872, 875

Nagar, D. K., Morán-Vásquez, R. A., and Gupta, A. K. (2015). "Extended matrix variate hypergeometric functions and matrix variate distributions." *International Journal of Mathematics and Mathematical Sciences*, 2015. MR3305177. doi: https://doi.org/10.1155/2015/190723. 895

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1): 3–48. MR1257793. 903

Pal, S. Sengupta, S., Mitra, R., and Banerjee, A. (2019). "Supplementary material: Conjugate Priors and Posterior Inference for the Matrix Langevin Distribution on the Stiefel Manifold." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1176SUPP. 874

Pearson, J. W., Olver, S., and Porter, M. A. (2017). "Numerical methods for the computation of the confluent and Gauss hypergeometric functions." *Numerical Algorithms*, 74(3): 821–866. MR3611557. doi: https://doi.org/10.1007/s11075-016-0173-0. 895

Rao, V., Lin, L., and Dunson, D. B. (2016). "Data augmentation for models based on rejection sampling." *Biometrika*, 103(2): 319–335. MR3509889. doi: https://doi.org/10.1093/biomet/asw005. 872

Schwartzman, A. (2006). "Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data." Ph.D. thesis, Stanford University. MR2708811. 876

Sei, T., Shibata, H., Takemura, A., Ohara, K., and Takayama, N. (2013). "Properties and applications of Fisher distribution on the rotation group." *Journal of Multivariate Analysis*, 116(Supplement C): 440–455. MR3049915. doi: https://doi.org/10.1016/j.jmva.2013.01.010. 873, 892

Turaga, P., Veeraraghavan, A., and Chellappa, R. (2008). "Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision." In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE. 871

van Dyk, D. A. and Meng, X.-L. (2001). "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics*, 10(1): 1–50. MR1936358. doi: https://doi.org/10.1198/10618600152418584. 872